Cooperative Radio Resource Management for Next Generation Systems

Mihovska, Albena D.

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Cooperative Radio Resource Management for Next Generation Systems

PhD work,

Albena Mihovska,

Center for TeleInfrastruktur, Aalborg University

September 2008

Supervisors: Prof Dr Ramjee Prasad
Co-supervisor: Dr Neeli Prasad

Abstract

This thesis proposes a novel framework for interactions belonging to entities that implement functionalities for radio resource management (RRM) in the scope of next generation systems.

Next generation systems will support a variety of heterogeneous networks, which means that connections span over different transport technologies supported by radio access networks of various topologies.

Access layer differences manifest themselves at radio link (e.g., delay and jitter) and at system level (radio resource allocation). This imposes the challenge of efficient network design and management. In this context RRM strategies are responsible for the efficient utilization of available resources in the radio access networks (RANs). Interworking between heterogeneous RANs is important for the provision of seamless mobility and ubiquitous coverage to users (i.e., user-perceived quality of service), for efficient network management including optimized system capacity, and fast deployment of new communication technologies.

Two main scenarios for interactions at layer 2 and layer 3 are considered, namely:

1. Inter-system interactions between RRM entities belonging to different RANs;

2. Intra-system interactions between RRM entities belonging to the same RAN and associated with different layer 1 transmission modes of the same radio access technology.

Existing RANs at this moment can be modified or updated for cooperation only at higher layers of the mobile network, which translates into routing at the radio network controller (RNC) or an equivalent network element. Further, already developed RRM architectures are optimized only for networks using a single layer technology. IMT-Advanced candidate RANs are foreseen of a simplified RAN architecture where the functionalities for RRM are moved closer to the air interface according to a distributed approach. This implies a large number of possible scenarios requiring resource allocation.

The heterogeneity of scenarios makes RRM interactions complex and multiple, meaning significant delays to execute them, degradation of QoS for mobile users, reduced throughput, unnecessary load increase in the networks and so forth. It is,

therefore, burdensome and inefficient to optimise traditional RRM mechanisms for a vast majority of specific scenarios.

The following are the key research contributions to resolve the above challenges:

- A novel, generic and scalable architecture for support of inter- and intra-system interworking described in terms of physical entities and logical functionalities. The entities in charge of decision making can execute reconfiguration actions to the underlying entities/network(s), by mechanisms capable to modify the network behavior. The network behavior is examined through simulations and key performance indicators (KPIs) real-time monitoring. Scalability is achieved by a combined centralized and distributed approach towards cooperation among architectural entities depending on the arisen situation and assessed in terms of achievable gains. The architecture is introduced in Chapter 1 and further assessed in relation to a variety of proposed RRM strategies in the rest of the Chapters of this thesis.
- Cooperation algorithms for mobility management, admission, load and congestion control in support of inter-and intra-system interworking towards provision of QoS to end users. Centralized and distributed strategies have been considered and a combined centralized and distributed approach has been proposed to provide for scalability. The proposed algorithms are generic and provide cooperation in any scenario without requiring major modifications in existing elements. They have been assessed for different traffic load scenarios and measurement strategies. The advantages of each algorithm have been assessed by simulations in terms of parameters related to QoS. Cooperative RRM algorithms are proposed and assessed in relation to different measurement strategies in Chapter 2. Chapter 3 proposes and assesses RRM strategies in support of inter-and intra-system handover. Chapter 4 proposes and assesses congestion, admission and load control in the context of inter-system interworking. Chapter 5 proposes a network-controlled mobility management scheme with policy enforcements. Chapter 6 proposes and assesses a multi-stage admission control strategy in support of intra-system interworking.
- Demonstration of the benefits of the proposed schemes by deploying them in a real-time simulation platform. The real-time simulation platform is a mobile IPv6-based software implementation where the IMT-Advanced candidate RAN is part of an experimental testbed configuration and the legacy RAN is an 802.11a/b/g WLAN access point. The efficiency of the proposed cooperative RRM is shown for a scenario of inter-system handover of a user with an ongoing real-time video streaming

application. Real-time network monitoring is based on use of key performance indicators, for which the process of decision making is enhanced by computational intelligence. The real-time simulation platform is proposed in Chapter 7.

The achieved results provide useful feedback relevant for the RRM design specifics to be considered in the context of next generation communication systems. The follow up research is intended to elevate the proposed concepts to a fully autonomic management framework. Focus is on cross-layer philosophy as a means for optimizing the mutual exchange of information between decision entities responsible for cooperative RRM.

Dansk Resume

Denne afhandling foreslår en nyskabende struktur for samspil mellem enheder, der implementerer funktionaliteter til administration af radioressourcer (RRM) indenfor næste generation kommunikationssystemer.

Næste generation kommunikationssystemer vil understøtte en bred vifte af heterogene netværk, hvilket betyder, at forbindelser mellem enheder fungerer over forskellige transportteknologier understøttet af radiotilgangsnetværk med forskellige topologier.

Forskelle i tilgangslag manifesterer sig på radioforbindelse- (f.eks. forsinkelse og jitter) og på system-niveau (radio ressourcefordeling). Dette medfører en udfordring i form af effektiv netværksudformning og administration. I denne sammenhæng er RRM strategier ansvarlige for en effektiv udnyttelse af de tilgængelige ressourcer i radio tilgangsnetværk (RANs). Samspil mellem heterogene RANs er vigtigt for levering af problemfri mobilitet og udbredt dækning for brugere (dvs. brugerens oplevelse af servicekvaliteten), og for effektiv administration af netværk, herunder optimeret systemkapacitet, og hurtig udbredelse af nye kommunikationsteknologier.

To vigtige scenarier for interaktioner på lag 2 og lag 3, behandles, nemlig:

1. Inter-system-samspillet mellem RRM enheder, der tilhører forskellige RANs;

2. Intra-system-samspillet mellem RRM enheder, der tilhører samme RAN og er forbundet med forskellige lag 1 transmissionsformer af samme radiotilgangsteknologi.

Eksisterende RANs kan udelukkende ændres eller opdateres for samarbejde på de højere lag af det mobile netværk, svarende til routing på radionetværkscontrolleren (RNC) eller et tilsvarende netværkselement. Yderligere er allerede udviklede RRM arkitekturer optimeret udelukkende til netværk, der benytter en enkelt-lags teknologi. IMT-Advanced kandidat-RANs forventes med en forenklet RAN arkitektur, hvor funktionerne for RRM er flyttet tættere på den trådløse grænseflade i henhold til en distribueret tilgang. Dette indebærer en lang række mulige scenarier, der kræver ressourcetildeling.

Forskelligheden i scenarier gør RRM-interaktioner komplekse og mangfoldige, hvilket betyder store forsinkelser ved at udføre dem, ledende til forringelse af servicekvaliteten for mobile brugere, reduceret kapacitet, unødvendig stigning af

trafikbelastning i netværkene osv. Det er derfor besværligt og ikke effektivt at optimere traditionelle RRM-mekanismer for langt størstedelen af specifikke scenarier.

Det følgende er de vigtigste forskningsbidrag til at løse de ovennævnte udfordringer:

• En ny, generisk og skalerbar arkitektur, der kan fungere på inter- og intra-system-niveau beskrevet i form af fysiske enheder og logisk funktionalitet. Enheder med ansvar for beslutningstagning kan udføre omkonfigureringsprocedurer til de underliggende enheder/netværk, ved hjælp af mekanismer, der er i stand til at ændre netværkets virkemåde. Netværkets opførsel er undersøgt gennem simulationer og realtidsovervågning af centrale ydelsesindikatorer (KPIs). Skalerbarhed er opnået ved en kombineret central og decentral tilgang til samarbejde mellem enheder i arkitekturen, afhængigt af den opståede situation og vurderet i forhold til de opnåelige gevinster. Arkitekturen er introduceret i kapitel 1 og yderligere vurderet i forhold til en række forslag til RRM-strategier i resten af kapitlerne i denne afhandling.

• Algoritmer til samarbejde for administration af mobilitet, optagelse, belastning og overbelastningskontrol til understøttelse af for inter- og intra-system-funktionalitet til bestemmelse af servicekvalitet for slutbrugerne. Disse er beskrevet enkeltvist som centraliserede og distribuerede strategier og en kombineret centraliseret og distribueret tilgang er foreslået for at opnå skalerbarhed. De foreslåede algoritmer er generiske og muliggør samarbejde i ethvert scenario, uden at større ændringer i eksisterende elementer er påkrævede. Algoritmerne er blevet vurderet i forskellige trafikbelastningsscenarier og med forskellige målestrategier. Fordelene ved hver algoritme er blevet vurderet ved simuleringer af parametre i forbindelse med servicekvalitet. Forslagene til kooperative RRM-algoritmer præsenteres og vurderes i forhold til forskellige målingsstrategier i kapitel 2. Kapitel 3 foreslår og vurderer RRM-strategier til understøttelse af skift mellem systemer på inter-og intra-system-niveauer. Kapitel 4 foreslår og vurderer overbelastnings-, optagelses- og belastningskontrol i forbindelse med inter-system kompatibilitetsproblemer. Kapitel 5 foreslår en netværkskontrolleret administration af mobilitet med regelbaserede håndhævelser. Kapitel 6 foreslår og vurderer en etapevis adgangskontrolstrategi til understøttelse af intra-system kompatibilitetsproblemer.

• Demonstration af fordelene ved de foreslåede ordninger er opnået ved at implementere dem i en realtidssimuleringsplatform. Realtidssimuleringsplatformen er en mobil IPv6-baseret softwareimplementering, hvor en IMT-Advanced kandidat-RAN er en del af en eksperimentel opsætning og den eksisterende RAN er et 802.11a/b/g WLAN-adgangspunkt. Effektiviteten af det foreslåede kooperative RRM er vist for et

scenarie af inter-system overdragelse af en bruger med en igangværende realtidsvideo streaming-applikation. Realtidsovervågning af netværket er baseret på anvendelse af centrale ydelsesindikatorer, med hvilke beslutningsprocessen er styrket med computerintelligens. Realtidssimuleringsplatformen præsenteres i kapitel 7.

De opnåede resultater giver nyttig feedback, der er relevant for design af RRM løsninger og som skal overvejes i forbindelse med næste generation af kommunikationssystemer. Opfølgende forskning er påtænkt for at bringe de foreslåede koncepter ind i en fuldt autonom administrationsstruktur. Fokus er på samarbejde på tværs af protokollag som et middel til at optimere den gensidige udveksling af oplysninger mellem beslutningsenheder, der er ansvarlige for kooperativ RRM.

Acknowledgements

First of all, I would like to thank my main supervisor, Prof. Dr. Ramjee Prasad who has supported me in my research path for many years and has inspired me to work with thoughtfulness and determination. Working with Prof Prasad has been a challenging and a rewarding experience that had taught me a lot not only about science and innovation but about life, in general, as well.

I also would like to thank my co-supervisor, Dr Neeli Prasad for giving me scientific guidance and moral support during the writing of this thesis.

I am much obliged and would like to thank Dr Jorge M. Pereira from the European Commission for the many fruitful scientific discussions and good advice during nine years of collaboration.

I would like to acknowledge the support, both scientific, and as a friend of Dr Sofoklis Kyriazakos who always found time to discuss and talk to me when it was needed.

I would like to thank my colleagues from the Center for TeleInfrastruktur (CTIF) and the Department of Electronic Systems at Aalborg University for the good working and social environment. In particular, I would like to thank my colleagues from the NETSEC group headed by Prof Ole Brun Madsen. Without the help of Rasmus Nielsen, Rasmus Olsen and Jimmy Nielsen, I would not have managed to put together the Danish translation of my PhD abstract.

I would like to thank my colleagues from the IST projects WINNER And WINNER II, and in particular, Dr Jijun Luo, Dr Elias Tragos, Emilio Mino, Dr Simon Plass, Dr George Karetsos, and Dr Seshaiah Ponnekanti.

Finally, I would like to thank my family and my friends for their strong support.

Table of Contents

Chapter 1

Introduction

This research work focuses on the interworking between heterogeneous radio access networks as a means for provision of quality of service (QoS) to mobile users and system capacity optimisation. The adopted scenario encompasses interworking between legacy and emerging mobile communication systems (i.e., IMT-Advanced candidate systems [1]-[5]). The objective was to design a set of specific strategies and algorithms for radio access networks (RANs) leading to an efficient use of available radio system resources for the support of services and mobile users over heterogeneous networks [6]-[20].

1.1 Problem Definition

This research proposes solutions for multi-link, multi-network radio resource provisioning and control. The main goal is to provide common and consistent RRM control in an integrated heterogeneous scenario where decisions must be made based on information from a multitude of parameters, some not directly mapped to each other and belonging to different technology domains and, where the heterogeneity of scenarios makes RRM interactions complex and multiple, meaning significant delays to execute them, degradation of QoS for mobile users, reduced throughput, unnecessary load increase in the networks and so forth.

Heterogeneity of networks means that connections span over several networks that deploy different radio access and transport technologies and that the networks are owned and operated by separate organizations. This indicates four specific cases of interaction, namely, *single technology-single domain, single technology-multi domain, multi technology-single domain, multi technology-multi domain.* With the introduction and integration of several systems with several modes and several layers, resource management becomes a more and more complicated task. For example, handover and load sharing algorithms must not only maintain the connection at a reasonable quality,

they should also consider whether it would be beneficial to move the connection to another system/layer/mode. This decision is not solely based on changing radio propagation, but also on system load, operator priorities and service quality parameters.

Current RRM solutions consider the single-technology-single-domain case, where radio resources are managed solely at the link layer (L2). With a single technology but multiple domains it is also possible to have a L2 solution. On the other hand in "native-IP" environment this could cause conflicts with the network layer (L3) interactions that will be taking place. Therefore communication with L3 entities is very important.

When multiple technologies are introduced, different link layers (L2) will interact with each other and there should be a layer, which will be the bridge between the technologies. This layer could be the IP layer (L3) acting through a dedicated interface. At L3 a decision can be made for the best resource management across the multiple technologies. In the multi technology-multi domain case, L3 decisions are needed not only in order to allow for cross-technology RRM, but also to remove any inter-domain management conflicts at L3.

Therefore, it is almost impossible to guarantee the promised QoS in a multi-technology and multi-operator environment. In case when the service does not meet the promised quality, it is equally difficult to identify the cause for that.

The key research issues and contributions of this thesis are proposed as solutions to the problem defined above. These are listed in Section 1.1.1 below.

1.1.1 Key Research Issues and Contributions

The key research issues towards a solution of the defined problem can be summarized as follows:

- Identify, propose, assess and validate RRM algorithms at layer 2 and layer 3 for an IMT-Advanced candidate system that supports multi-hop communications and is used as the reference system (intra-system interworking). The RRM algorithms relate to mobility management, congestion, admission and load control of IMT-Advanced mobile users. Further, these consider the state of the transport network during the decision making process;
- Identify, propose and validate cooperative RRM mechanisms and strategies for inter-system interworking between legacy systems (e.g., GPRS, UMTS, WLAN) and IMT-Advanced systems. The cooperative RRM mechanisms are based on

the common RRM (CRRM) [21]-[23] and joint RRM (JRRM) [24]-[25] principle and assume tight coupling [26] between the networks.

- Provide a flexible, generic and scalable framework for support of the cooperative RRM mechanisms that does not require changes in the RRM architectures of existing RANs and ensures successful interworking in any scenario. The framework increases the performance of network segments related to RRM and implements inter- and intra-system cooperation at various levels of the RAN architecture to ensure effective management of radio resources as a prerogative to the provision of seamless mobility, ubiquitous coverage, efficient network management and fast network deployment in next generation mobile communication systems.
- Demonstrate the benefits of the proposed mechanisms and framework in terms of improved performance. This is achieved through an implementation in a real-time simulation platform based on Mobile IPv6 and by means of a real-time video streaming application.

1.1.2 Technical Approach

It is proposed here that cooperation (i.e., RRM) mechanisms for intra-system interworking are developed at the radio segment level of the new RANs following a combined centralized and distributed approach, while inter-system interworking is handled by a central entity located outside of the RAN architecture following a centralized approach. The technical approach to the defined problem is shown in Figure 1-1.

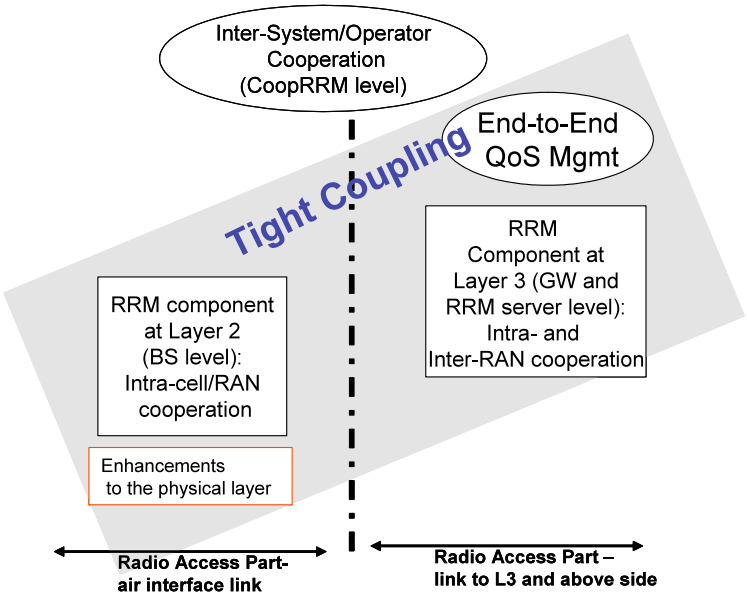


Figure 1-1 The proposed technical approach to Cooperative RRM

Intra-system RRM mechanisms are implemented at network elements of the rank of base stations (BSs), relay nodes (RNs) and below. It is possible, however, to activate cooperation mechanisms from the network side of the radio access level to optimise the overall network and radio performance [i.e., gateway (GW)]. Additionally, other elements should also be able to activate them to ensure, for example, end-to-end requirements for QoS.

Inter-system interworking mechanisms are activated by an entity referred to as CoopRRM entity, located outside of the RAN. The GW entity is an anchor point for inter-system cooperation and provides the interface towards the CoopRRM entity, the Internet, and the operator services.

The adopted technical approach comprises the following steps:

- Identification, definition and assessment of RRM algorithms for mobility management, congestion, admission and load control for inter-system and intra-system cooperation. Focus was on defining mechanisms that are common for all user scenarios and applications defined for next generation systems [1], [3] rather than optimizing per type of algorithm and scenario. The proposed algorithms were assessed for different mobility scenarios (macro-cell, micro-cell, indoor), for a service mix (e.g., conversational, interactive, streaming, background) and for different traffic load scenarios (normal, busy, emergency/sports event). Results show an improvement of delays, mean user throughput, of system capacity for higher loads, and blocking and dropping probabilities compared to state of the art proposals. Decision making for the activation of a cooperative RRM includes use of triggers and KPIs calculation/monitoring. For the calculation of the KPIs a special reward function has been defined. The results of this step of the research are included in Chapter 2, Chapter 3, Chapter 4, partly, and Chapter 6.
- Investigation of the impact of the proposed RRM algorithms on the proposed architecture for cooperation in terms of scalability and flexibility. The resulting architecture is described in terms of logical and physical functionalities. The following functionalities have been implemented for the cooperative RRM:
    - Functionality for mobility management (e.g., handover);
    - Functionality for congestion, admission and load control;
    - Functionality for QoS management.

The proposed RRM algorithms are investigated in further details for the resulting architecture and assessed in terms of signaling load for a centralized and distributed approach to RRM. The advantages of a joint approach are presented in terms of achievable gains. The results of this step of the investigation are included in Chapter 4.

- Investigation of the benefits of the proposed interworking strategies for mobility management and load and admission control in an IMT-Advanced candidate system suitable for multi-hop communications. The results of this investigation include a novel proposal for policy-based mobility management (e.g., handover) framework mechanism that manages policies related to flow control and user context transfer during handover and RAT association, (in Chapter 4) and a novel multi-stage admission control based on load dependent decision polling that takes into account the available resources both in the RAN and the backbone network (in Chapter 5). The results show improvements in terms of reduced delays (for user context transfer during radio handover) and achievable gains through load balancing and reduced response times (for the multi-stage admission control).
- Validation of the proposed RRM strategies through a heterogeneous realization in a real-time simulation platform and for a real-time video streaming application. The KPI reward function and decision making process for the activation of an RRM algorithm are further investigated. An approach for efficient decision making during handover and admission control is proposed based on use of computational intelligence (i.e, fuzzy logic). Results are based on real-time testing and are in terms of improved performance. This research step and the results are included in Chapter 6.
- Outline of future work as a follow up of the performed research. Future work envisions short-term and long-term investigation work towards a completely autonomous framework for RRM in the adopted scenarios. Short –term research will focus on the role of the BS-BS interface (i.e., wireless versus wired interfaces) and its impact on the efficiency of user context transfer during handover, the development of the complete mathematical framework for the multi-stage admission control that would improve the reported here assessment results. Long-term research includes the further development and implementation of the decision making framework based on computational

intelligence combined with cross-layer design for achieving global QoS in next generation communication networks. Future work is described in Chapter 7.

The outlined research topics were addressed by the proposed framework for cooperative RRM, which is aligned as much as possible with the proposals envisaged by 3GPPP and the ITU for next generation communication systems.

1.2 Motivation for the Carried Out Research

This research work was motivated by the trends of convergence and interoperability envisioned for current and next generation mobile communication systems [1] and implied by the variety of radio access technologies (RATs) emerging because of the rising demand for fast, scalable, efficient and robust data transfer over the air. The following challenges can be identified:

- Heterogeneity in ownership, technologies and applications implying the need for the following:
  - Coexistence and interworking between legacy, evolving and newly emerging communication systems and technologies while allowing for self-contained individual systems so that benefits from previous investements in legacy technologies can be further exploited;
  - Coexistence and interworking between different PHY layer modes of the same radio system [2] for ubiquitous coverage;
  - Establishment and the maintenance of connections with required quality under various scenarios;
  - Reducing the infrastructure costs but achieving higher performance leading to new types of network nodes.
  - Optimised control, management and flexibility of the future network infrastructure.
- User-centricity of the communication process requiring resource allocation based on individual and specific needs but achieving the following:
  - Uninterrupted coverage even in remote areas and for any application and mobility scenario;
  - Global roaming capabilities implying also the need for new business models;
  - Fast scalable, and efficient access to system services.
- Separation between the underlying networked infrastructure and the services/applications that requires the following:

- Simultaneous use of various transmission technologies for the delivery of the same service;
- Cooperation at vertical level as well as horizontal (i.e., across architectures and providers) to allow for the Internet Service Provider (ISP) to guarantee the promised quality in a multi-technology and multi-operator environment.

1.3 Background

The strong demand in wireless systems, including broadband, requires more capacity of advanced mobile communication systems and promotes mixed solutions depending on the capacity and coverage area required for a certain service.

Next generation communication systems are seen as organised in a layered structure, comprising of a distribution layer, a cellular layer, a hot spot layer, a personal network layer and a fixed (wired) layer [3]. These systems shall also support the use of coverage enhancing technologies [3]. However, legacy systems, such as GSM, UMTS, and WLAN, would continue to provide services to users. Therefore, a generic framework for the support of the interworking between these, essentially, different systems is required.

Interworking between WLAN and UMTS networks has been a research topic driven primarily by ETSI / BRAN [26] and 3GPP [4]. The feasibility of UMTS and WLAN interworking was drafted in the recommendation 3GPP TR 22.934 [27], where not only different levels of interworking but also different environments were defined. Broadly, it was classified as *loose coupling* and *tight coupling*. From a macro point of view the main difference is how and where one RAN is coupled to another network. The choice is mainly a trade-off between the required degree of modifications to standards, the seamless degree of interworking and the complexity of the common infrastructure.

Substantial effort has been put into the design of cooperation architectures supporting the interworking among heterogeneous communication systems. Work has been performed in standardization groups [26], [27], and in various EU-funded IST projects on architectures and platforms for cooperation schemes between heterogeneous RANs [28], [29], [31], [32]. However, the research essentially focused on cooperation between UMTS, GSM/GPRS, and WLAN [33]-[37] or between legacy and evolving from UMTS networks (i.e., E-UTRAN [5]) [32]. In [38], the gains of a joint RRM approach were investigated. This concept was investigated further within the IST Project E2R [39] where focus was on coexistence of RATs within the user terminal

(terminal reconfigurability). The reconfiguration technology provides adaptation of the radio interface to varying RATs, provision of possible applications and services, update of software and enabling full exploitation of flexible resources and services of heterogeneous networks. Reconfigurable terminals, with embedded radio link layer functionalities according to network architecture, will be able to enable cooperation between multiple RATs. The IST project AROMA [32] provided a proposal and a detailed analysis of algorithms for admission control and cell selection in a heterogeneous scenario, also taking into account the available resources in the transport network.

In the scope of next generation communications research, the IST project WINNER [2] proposed an IMT-Advanced [1] candidate system based on the OFDM technology whose system comprises various functions that are intended to avoid data loss and minimize delay during handover, as well as ensure coverage in the recommended by the ITU deployment scenarios [3]. This covers rural areas and provides also a contiguous coverage layer in towns and cities, where it will overlap with metropolitan and local area deployments. Another important target is to support the full range of mobility scenarios up to high speed trains. One requirement is that a ubiquitous radio system has to be self-contained, allowing it to target the chosen requirements without the need for interworking with other systems. Another requirement is that cooperation, whenever required, will be successfully ensured between any new or legacy systems.

The proposed RAN architecture is simple and supports a distributed approach to functionalities implementation similar to the one adopted by the Third Generation Partnership Project (3GPP) [4] for the design of the Long Term Evolution (LTE) [5] systems.

Following the preliminary trends in the 3GPP architecture evolution [4] it is noted that the RAN is moving towards an open distributed topology. This is relevant for the underlying radio level activation of the resource management regime [6], [42].

1.4 Research Scenario

The reference scenario for the performed research is shown in Figure 1-2. The scenario shows two cases of mobility of a user: 1) from an IMT-Advanced candidate system to an UMTS or WLAN system, and vice versa, which is requires inter-system cooperation and 2) within the IMT-Advanced system, between cells served by different PHY –layer modes, which requires intra-system cooperation.

The IMT-Advanced candidate RAN (also assumed as the reference RAN in this research work) consists of the following entities: GW, BSs, fixed and mobile RNs (FRNs and MRNs, respectively). The RNs are used to extend the coverage of a BS or to give coverage to shadowed zones (i.e., a multi-hop communication system).

The protocol stack in the RN is the same as in the BS. As compared to systems of previous generations, IMT-Advanced candidate systems have moved the radio-related functionality closer to the radio interface. The radio interface is evolving towards flexible architecture and is designed to operate efficiently for different deployment modes spanning over a ubiquitous coverage, such as local area (LA), metropolitan area (MA) and wide area (WA) [2].

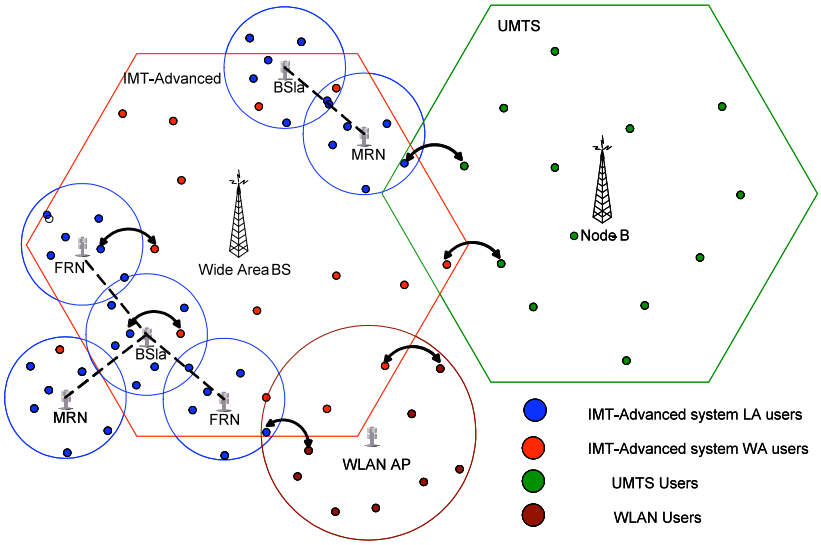


Figure 1-2 Reference research scenario.

It is proposed that inter-system interworking for the scenario in Figure 1-2 is based on the tight coupling principle (i.e., the external entity in charge of inter-system interworking will be involved in each RRM decision). Tight coupling is also the assumption throughout the thesis.

In future heterogeneous wireless networks, RRM must be coordinated across a number of access technologies coexisting within the same network. Inter-RRM signalling is also required in order to transfer the information between RRM entities upon which resource allocation and admission control decisions can be based.

1.5 Reference Architecture for Cooperative RRM

The basic reference framework for cooperation based on the reference scenario in Figure 1-2 is shown in Figure 1-3.

The architecture in Figure 1-3 supports a centralised and distributed approach to RRM. A centralised approach was proposed for the inter-system cooperation [40], [41], for which the main decision making point is the CoopRRM entity located outside of the RANs. One requirement of the cooperation architecture is to provide some inter-RAN services such as admission control, handover, scheduling, and QoS-based management, and other services, such as billing, authentication, authorization. Tight coupling was selected as the most suitable degree of coupling between wireless networks and entities of the same RAN (e.g., BSs) [42]-[44].

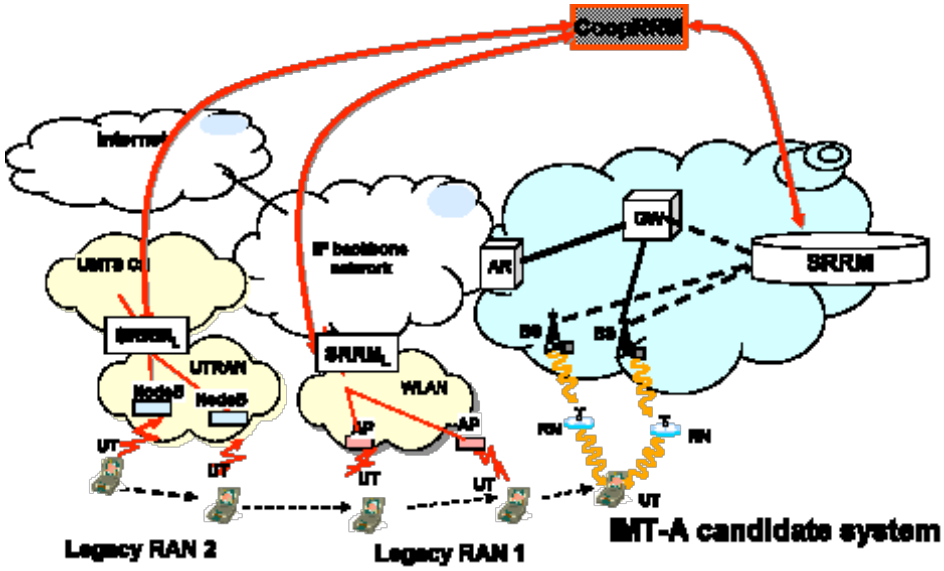


Figure 1-3 Basic reference framework for cooperative RRM.

A specific RRM entity (SRRM) implementing the RAN- specific RRM mechanisms is located within each RAN. In a situation, when a local RRM approach is not sufficient to ensure seamless user mobility, the decision center would be shifted to the CoopRRM that would execute and appropriate algorithm to resolve the occurred problem. Thus, the proposed implementation is of generic nature that does not require major changes in the individual RAN architectures, and allows for easy inclusion of any newly designed RAN.

The SRRM module located in the legacy RAN (hereafter referred to as $SRRM_L$) implements two types of functionalities and interfaces, one for **traffic monitoring** and reporting of physical legacy nodes and the other devoted to the **direct actuation** of the RRM algorithms in the legacy RAN nodes. In other words, it translates the CoopRRM commands to the legacy RAN. The SRRM in the IMT-Advanced reference RAN (hereafter referred to as $SRRM_W$) implements the monitoring and actuation

functionalities and also the support of the functionalities related to the inter-RAN cooperation and the internal RRM coordination functionality (i.e., $SRRM_W$ is distributed in the RRM server, GW and BS, respectively). This two-way communication is for transferring monitoring information, but also for executing global RRM techniques. For the congestion case, the CoopRRM will not be able to change any of the parameters of a legacy RAN, but for handover cases the CoopRRM could change some RRM parameters in the legacy RANs. For the reference protocol architecture in Figure 1-4 (e.g., it shows cooperation between the IMT-A reference RAN and UTRAN) the arrow pointing to the SRRM of the legacy RAN shows the indirect interaction as a result from the shift (handover) of users from the IMT-A reference RAN to the legacy RAN. For example, when the CoopRRM decides for inter-system handover of one user to UMTS, after having checked its status, the user will request a radio resource control (RRC) session establishment, which is, in this case, the indirect interaction of the CoopRRM with the UMTS SRRM. The CoopRRM will also have interfaces with other CoopRRM of the same or different operators.

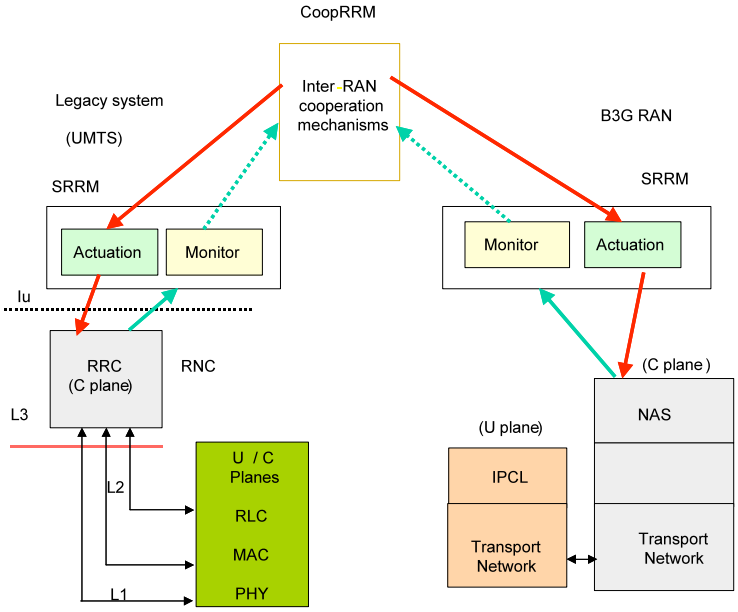


Figure 1-4 Reference protocol architecture for cooperative RRM

There are two possibilities for the inter-system cooperation. For example, in the case of mobility management, the CoopRRM either can advise the SRRM entities only before the decision, or the CoopRRM decisions are binding for the SRRM entities.

This means that inter-system cooperation may be realized also at lower layers, in consequence, the SRRM will be associated to or will reside in the corresponding entities of the legacy RANs, (i.e., RNC, BTS and MG of UMTS, GSM and IEEE802.11 networks, respectively.) For the IMT-A reference RAN, inter-system cooperation

functionalities will reside in the GW, which is split into logical functionalities related to the user plane (GW_{UP}), composed of those protocols that implement data transfer services of the actual user-data, and control plane (GW_{CP}), composed of protocols for control of transfer services and connections between the access network and the UT. The protocol termination for the user and control planes in the GW in support to inter-system cooperation is shown in Figure 1-4. The IP convergence layer (IPCL) is the protocol proposed for the user plane (UP) and Non-Access Stratum (NAS) is the protocol for the control plane (CP) [45]. The NAS is a functional layer in the protocol stack that supports signalling and traffic between the core network and the UT [47].

Normally, the SRRM will implement the functionality to translate RRM messages between the CoopRRM and the UT. The B3G monitor set of the $SRRM_W$ will include the legacy RANs cells, and the legacy RAN monitor set of the $SRRM_L$ will include the cells for the B3G RAN. The monitoring functionality will initiate a request of actuation to the CoopRRM entity, when a trigger is activated by a measurement that shows that a threshold is surpassed.

The handover procedure in this case will be similar to the one used for UMTS–GSM handover, with messages transmitted by the RANs, for example, *[HANDOVER FROM UTRAN COMMAND]*; this command sends an encapsulated message of the other RAN, that contains all the needed information to allow the UT to connect to the RAN. It is necessary to update the messages and protocols of the legacy RAN to include IMT-A RAN messages (and perhaps other legacy RANs). In this case the SRRM would use the monitoring functionality and should interact with the other SRRM entities to transmit messages in the other RANs.

The cooperative RRM mechanisms presented in this thesis exploit two types of RRM schemes proposed for interworking, namely:

- Common Radio Resource Management (CRRM) defined within 3GPP to allow better inter-working between UMTS and GSM/GPRS networks [4], [21], [22], [23], [43], [44].
- Joint Radio Resource Management (JRRM) as defined in [24], [25], [38], [39] for inter-working between WLAN (e.g., HIPERLAN2) and UMTS.

The cooperative RRM framework presented in this thesis adopts for inter-system interworking the CRRM model (i.e., a central entity is in charge of RRM decisions) and for intra-system interworking, the JRRM model (i.e., a central and internal to the RAN entity manages the overall capacity of the basic physical nodes in situations of high loads and joint management of traffic streams between entities). Some strategies

adopted by the Concurrent RRM [26] approach have also been considered for a totally distributed RRM execution.

The main benefits of evolving the existing approaches are that optimal system performance can be achieved with limited changes and already existing functionalities.

The location of the RRM functions can be divided between the link layer and the network layer, considering the information requirements and functions that are available at other layers. The division of the RRM architecture on each layer is based on the "target object" or "environment" that will need the RRM function. However, there are some cases, where the RRM entity is relevant in both layers (L2 and L3). In these cases, the function is divided across both layers with different aspects of the function resident in different places coinciding with different "target objects" or "environment". For functions of that kind (that split between two layers) there must be close cooperation between layers to ensure efficient RRM control.

Benefits of a centralised RRM are achieved at the expense of a higher computational complexity since a larger interchange of information among network agents increases the signalling. Delay in signalling is higher than in the distributed approach, but the reaction time is not as critical because of the vertical handover and because the legacy RAN functionalities are slower than the radio functionalities envisioned for IMT-Advanced candidate systems.

It is proposed here that the following entities are implemented to realise the functionalities for mobility management, congestion, admission and load control, and for QoS management, namely:

- Handover decision entity for making the final decision regarding the target RAT for the UT to handover;
- Triggers entity for collecting/comparing triggers and deciding whether a handover process has to be initiated;
- Measurements entity for collecting measurements from the current and/or other RATs/modes (periodically) and calculating extra values;
- RATs monitoring and filtering entity to keep track of the available RATs/modes as well as keeping a list on the available RATs/modes that each UT can access based on user preferences, network operator restrictions, UT capabilities;
- User preferences entity for keeping track of the user context information such as cost, RAT preferences, QoS classes;
- A central admission control entity, which will be responsible for the final decision and located in the CoopRRM or RRM server;

- A local admission control entity (in each SRRM) cooperating with the measurement entity in each RAN, and checking the different network admission criteria, cooperating with the admission control entity in the CoopRRM (receiving/sending information) for selecting the ongoing sessions that will perform intersystem handover or will degrade their QoS in order to gain the needed resource for admitting the target session and for cooperating with other possible entities located in the SRRM (i.e. handover entity or QoS management entity;
- Entity responsible for maintaining the handover queue for maintaining a queue for the handover sessions that cannot be completed immediately and must remain in the queue until the needed resources become available;
- A network and session manager located at the CoopRRM for prioritising the sessions according to their service class and for decisions for assigning them to a network that would maintain the QoS requirements of the users. The main function of this entity is to build up and maintain an active set of candidate RANs available based on the user request and the relevant user profile;
- A multi-RAN scheduler located at the GW and RRM server to forward the packets within the cooperative RAN cluster to one or a set of candidate RANs depending on the bearer/service attributes. Further, tight coupling is used to map the output of the scheduler in charge of link adaptation to the multi-RAN scheduler.

The impact of the proposed cooperation architecture on the RAN is examined in detail in Chapter 2.

Figure 1-5 is a visual interpretation of the defined in Section 1.1 problem and shows the interdependencies between the different contributions proposed in this thesis. The thesis proposes a novel framework for RRM that ensures that the given radio resources are utilised at different layers of the radio access network (RAN) architecture. The individual contributions proposed in the scope of this work represent the various chapters of the thesis structure. The organization of this thesis is based on the interdependencies of the blocks part of the proposed RRM frameworkand is shown in Figure 1-6.

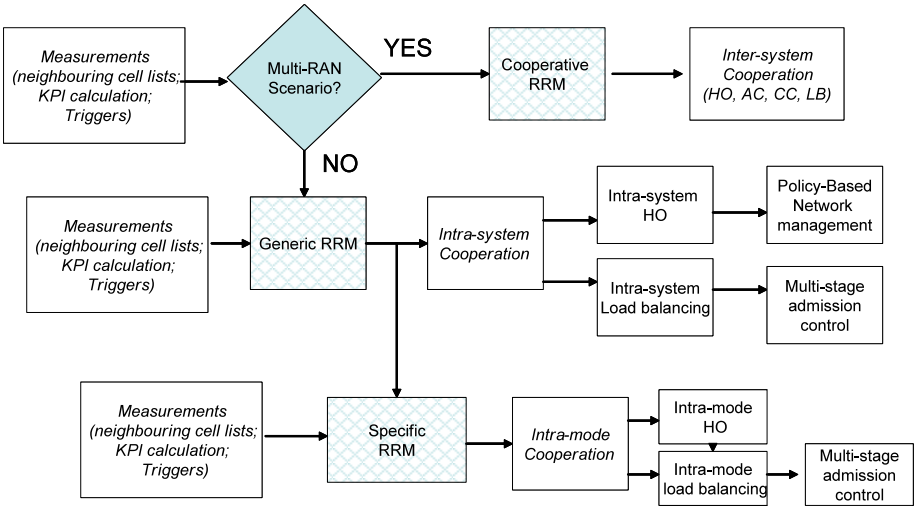


Figure 1-5 Topical interdependencies of the proposed RRM framework and contributions.

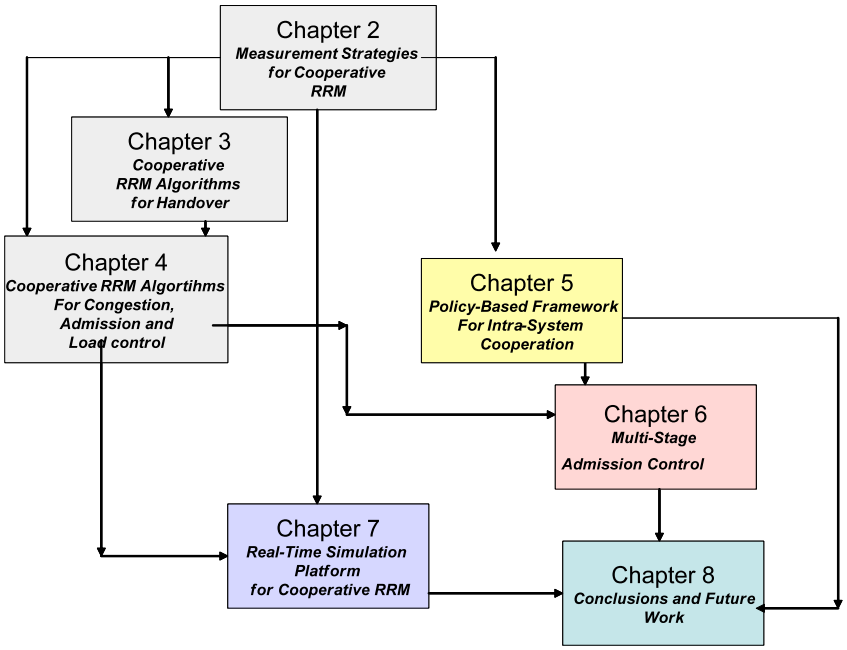


Figure 1-6 Interdependencies between the Chapters of this thesis.

1.6 Preview of This Thesis

This thesis is organized as follows.

Chapter 2 proposes measurement strategies for cooperative RRM strategies (i.e. inter- and intra-system cooperation). Measurements are obtained by (real-time) monitoring. Strategies, such as neighboring cell lists and key performance indicators (KPIs) can be used as triggers that would activate an RRM mechanism depending on the scenarios (see Figure 1-5), which are proposed in Section 2.1. Measurement strategies are proposed and evaluated in Section 2.2. The importance of triggers for the actuation of an RRM algorithm is also analysed. Measurements in a single-RAN scenario will be used to activate a generic or specific RRM in support of intra-system and intra-mode cooperation, respectively. Measurements are assessed for handover strategies (e.g., protocols and algorithms) proposed for inter- and intra-system cooperation in Section 2.3. Section 2.4 analyses the impact of the proposed schemes on the RAN architecture. Section 2.5 concludes the Chapter.

Chapter 3 proposes and evaluates RRM algorithms for inter-and intra-system handover. Inter-system handover is analysed in Section 3.1 for a scenario assuming an IMT-Advanced candidate system and a UMTS system. Intra-system handover is analysed in Section 3.2. In particular an optimised intraction between RRM functions is proposed based on a hybrid approach. Section 3.3 concludes the Chapter.

Chapter 4 proposes and assesses protocols and algorithms for inter-system admission, congestion and load control (see Sections 4.1 and Sections 4.2). An assessment for the performance of the proposed algorithms is given in terms of QoS related parameters in Section 4.3. A feasibility study for the cooperation impact on the radio layer is also given. Section 4.4 concludes the chapter.

Chapter 5 proposes a joint centralised and distributed approach to cooperation and proposes a hybrid approach as a means to achieve scalability in the cooperative RRM framework (see Section 5.1). Section 5.2 assesses the proposed scalable architecture in terms of achievable gains and signalling. Here, the multiplexing gain is introduced as a benefit from interworking between BS. The scalable architecture allows for a network-controlled policy-based approach to cooperative RRM. Section 5.3 proposes a novel network-controlled policy-based approach to mobility management, focused on policies related to RAT association during initial access and user context transfer during radio and IP handover. Section 5.4 concludes Chapter 5.

Chapter 6 proposes a novel approach for admission control in an IMT-A candidate system supporting multi-hop communications. The proposed mechanism is based on load dependent decision polling and takes into account the load in various parts of the

RAN as well as available resources of the backbone network. The scheme is investigated in terms of achievable gains from load balancing and BS interworking in Section 6.2 and an implementation is proposed and described in Section 6.3. Section 6.4 concludes the Chapter.

Chapter 7 proposes an implementation of the cooperative RRM architecture and realises this implementation as real-time simulation platform based on Mobile IPv6. Section 7.1 gives the motivation for the chosen implementation approach. Section 7.2 describes the requirements of the chosen implementation. Section 7.3 describes in further detail the importance of KPIs for the activation of a cooperative RRM and proposes the implementation of the monitoring process. Section 7.4 describes the different modules of the real-time simulation platform and their functionalities. Section 7.5 proposes an implementation of fuzzy logic as an enhancement of the decision making process during handover and admission control. The platform is realised both as a stand-alone implementation and as an integrated trial where the IMT-A candidate air interface is part of a test-bed configuration. This implementation is proposed and described in Section 7.6. The proposed algorithms for inter-system handover (Chapter 3) are assessed in this implementation and an improved performance is shown for a real-time video streaming application. Section 7.7 concludes the Chapter.

Chapter 8 concludes the thesis and proposes future short-and long-term research based on the achieved results.

References:

[1] International Telecommunications Union, ITU, at www.itu.int.

[2] IST Project WINNER and WINNER II, www.ist-winner.org

[3] RECOMMENDATION ITU-R M.1645, "Framework and Overall Objectives of the Future Development of IMT 2000 and Systems Beyond IMT 2000," At www.itu.int.

[4] Third Generation Partnership Project, 3GPP at http://www.3gpp.org

[5] Long Term Evolution, http://www.3gpp.org/Highlights/LTE/LTE.htm

[6] A. Mihovska, et al., "A Novel Flexible Technology for Intelligent Base Station Architecture Support for 4G Systems," *Proc. of WPMC'02,* Honolulu, Hawaii, October 2002.

[7] A. Mihovska, et al, "Towards the Wireless 2010 Vision: A Technology Roadmap," in *Special Issue on Advances in Wireless Communications of the Springer International Journal on Wireless Communications,* DOI: 10.1007/s11277-006-9180-0, September 2006.

[8] A. Mihovska and R. Prasad, "Secure Personal Networks for IMT-Advanced Connectivity," in *Special Issue of the Springer International Journal on Wireless Communications,* DOI: 10.1007/s11277-008-9485-2, April 2008.

[9] A. Mihovska, S. Ponnekanti, and R. Prasad, "Ensuring End-to-End QoS Through Dynamically Adaptive RRM Techniques," In *Proc Of WPMC'03,* October 2003, Yokosuka, Japan.

[10] A., Mihovska, et al., "Requirements and Algorithms for Cooperation of Heterogeneous Networks," in *Springer International Journal On Wireless Personal Communications,* DOI: 10.1007/s11277-008-9586-y, September 2008.

[11] A., Mihovska; H., Laitinen, P., Eggers, "Location and Time Aware Multi-System Mobile Network," in *Proc. of Mobile Location Workshop'03,* Aalborg, Denmark, May 2003.

[12] A. Mihovska, G. Karetsos, S. Ponnekanti, "RRM Techniques for Heterogenous Wireless Systems," in *Proc. of Mobile Venue Workshop,* May 2004, Athens, Greece.

[13] A., Mihovska, S. Kyriazakos, E. Gkroutsiosis and J. M. Pereira, "QoS Management in Heterogeneous Environments," in *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

[14] A. Mihovska, S. Kyriazakos, E. Mino, M. Pischella, E. Tragos, V. Sdralia, "Assessment of RRM Schemes for the Efficient Cooperation of RANs: WINNER Requirements," *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

[15] A. Mihovska, S. Kyriazakos, E. Mino, M. Pischella, E. Tragos, V. Sdralia, "Assessment of Radio Resource Management Schemes for Efficient Cooperation of RANs: An Implementation Approach," in *Proc. of IST EVEREST Workshop,* November 2005, Barcelona, Spain.

[16] A. Mihovska, S. Kyriazakos, and J. M. Pereira, "Algorithms for QoS Management in Heterogeneous Environments," *Proc. of WPMC'06,* San Diego, California, September 2006.

[17] A. Mihovska, J. Luo, E. Mino, E. Tragos, C. Mensing, G. Vivier, R. Fracchia, "Policy-Based Mobility Management for Next generation Systems," *Proc. of IST Mobile Summit 2007,* Budapest, Hungary, July 2007.

[18] A. Mihovska, S. Kyriazakos, and N. Prasad, "A Cognitive Approach to Network Monitoring in Heterogeneous Environments," in *Proc. of WPMC'07* ,December 2007, Jaipur, India.

[19] A. Mihovska, J. Luo, B. Anggorojati, S. Kyriazakos, N. Prasad, "Multi-Stage Admission Control for Load Balancing," in *Proc. of WPMC'08,* Sept 8-11, 2008, Lapland, Finland.

[20] A. Mihovska, E. Tragos, S. Kyriazakos, P. Anggraeni, N Prasad, "A Practical Implementation of Cooperative Radio Resource Management," in *Proc. of the ATSMA International Networking and Electronic Commerce Research Conference (NAEC 2008),* September 25-28, 2008, in Riva del Garda, Italy.

[21] Release 99, www.3gpp.org/Releases/3GPP_R99-contents.doc

[22] www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_26/Docs/PDF/SP-040900.pdf

[23] F., Meago, "Common Radio Resource Management (CRRM)", *COST273,* May 2002.

[24] J., Luo, R., Mukerjee, M., Dillinger, E., Mohyeldin, E., Schulz, "Investigation of Radio Resource Scheduling in WLANs Coupled with 3G Cellular Network," *IEE Communications Magazine,* June 2003, pp.108-115.

[25] J., Luo, E., Mohyeldin, N., Motte, and M., Dillinger, "Performance Investigations of ARMH in a Reconfigurable Environment," *SCOUT workshop,* Paris, 2003.

[26] ETSI TR 101 957: "Broadband Radio Access Networks (BRAN); HIPERLAN Type2; Requirements and Architectures for Interworking between HIPERLAN/2 and 3rd Generation Cellular Systems", V1.1.1 (2001-08).

[27] 3GPP TR 22.934, V1.0.0 Feasibility study on 3GPP system to Wireless Local Area Network (WLAN) Interworking Rel-6.

[28] Evolving Systems Beyond 3G, IST-2000-28584 MIND at http://www.ist-mind.org/.

[29] IST Project CAUTION, "Capacity Utilization in Cellular Networks of Present and Future Generation," at www.telecom.ece.ntua.gr/CautionPlus/.

[30] Advanced Radio Resource Management for Wireless Services, IST Project ARROWS, at http://www.arrows-ist.upc.es/.

[31] Evolutionary Strategies for RRM, IST Project EVEREST, at www.everest-ist.upc.es.

[32] Advanced Resource Management Solutions for Future All IP Heterogeneous Mobile Radio Environments, IST Project AROMA at http://www.aroma-ist.upc.edu/.

[33] UMTS, at http://www.umts-forum.org.

[34] HIPERLAN standards, at http:// www.etsi.org/.

[35] WLAN at http://en.wikipedia.org/wiki/Wireless_LAN.

[36] GPRS technology at www.gsmworld.com/technology/gprs/index.shtml.

[37] IEEE 802.11, The Working Group Setting the Standards for Wireless LANs, at http://www.ieee802.org/11.

[38] IST Project SCOUT, "Smart User-Centric Commuunication Environment," at http://www.ist-scout.org/.

[39] IST project E2R, End-to-End reconfigurability, http:// e2r.motlabs.com/

[40] M., Lott, et al., "Cooperation of 4G Radio Networks with Legacy Systems," Proc. of IST Mobile Summit 2005, Dresden, Germany, June 2005.

[41] A., Mihovska, et al., "Assessment of Radio Resource Management Schemes for Efficient Cooperation of RANs," Proc. of WPMC'05, Aalborg, Denmark, September 2005.

[42] E., Mino, A. Mihovska, et al., "Scalable and Hybrid Radio Resource Management for Future Wireless Networks," Proc. of IST Mobile Summit 07, Budapest, Hungary, July 2007.

[43] UTRAN Radio Resource Control Protocol Specification , TS 25.331, V 5.6.0, at www.3gpp.org, September 2003.

[44] UTRAN Radio Interface Protocol Architecture, Release 5, TS 25.301, V 5.2.0, www.3gpp.org, September 2002.

[45] A. Mihovska, et al. " Practical Implementation of Cooperative Radio Resource Management, " accepted for publication in *Proc. Of the 2008 Networking and Electronic Commerce Research Conference (NAEC 2008),* to be held in September 25-28, 2008, in Riva del Garda, Italy.

[46] A. Mihovska, et al., "D4.8.2: Cooperation Schemes Validation," Deliverable 4.8.2, IST project WINNER II at www.ist-winner.org.

[47] IST project WINNER II, Deliverable 6.13.14, "WINNER II System Project Description," November 2007.

Chapter 2

Measurement Strategies for Cooperative RRM

This Chapter proposes the measurement strategies for the cooperative RRM framework and evaluates it for RRM algorithms proposed here for inter-and intra-system handover as part of mobility management for IMT-A mobile users. Handover here is understood as the switching process between two radio systems (inter-system handover) or between two cells of the same/different mode/radio access system (intra-system handover). Further, the proposed RRM framework is applicable to inter-system interworking regardless of the type of the involved systems. The measurement strategies and proposed handover algorithms are assessed for different scenarios.

The protocol part of the proposed RRM algorithms includes specifications of the messages exchanged between the involved entities, the frequency of these messages and the associated interfaces.

The RRM functions rely on measurements performed by the RRM specific protocols. Measurements can be performed in a different way and they have to be reported in such a way that comparison of the obtained values is possible. The following measurements strategies are proposed in support of the RRM framework:

- Key performance indicators (KPIs);
- Neighbouring cell lists;
- Triggers;
- Measurements from use of location information.

Chapter 2 investigates what measurements should be considered with the cooperative RRM framework proposed in this thesis, and proposes what quantities should be measured, how frequently, and how measurements should be reported in order to trigger a given RRM algorithm. The proposed measurement strategies are assessed proposed here handover protocols and algorithms.

Further to the realization of RRM techniques, the location of the RRM functions is also studied. A combined centralised and distributed approach is proposed. Therefore,

the proposed RRM algorithms are made consistent with the specifics of the RAN architectures of the investigated IMT-A reference system and the RANs of the legacy systems. The location of the RRM functions within the network architecture is an essential issue and can affect the performance if causing significant signalling and delays. In a centralised architecture, a central entity monitors and makes decisions regarding the allocation of resources and the user terminal (UT) has a minimal participation. In a distributed RRM architecture, the decision entities for each RRM function are located to different nodes, including the UT. A hybrid approach is also proposed, and there the decision levels of the same RRM functionality that can be active at different timescales are allocated to different nodes. The impact of the proposed algorithms on the proposed cooperation architecture is also studied in Chapter 2.

Chapter 2 is organised as follows. Section 2.1 defines the scenarios for inter- and intra-system cooperation. Section 2.2 defines, proposes, and analyses measurements strategies for the actuation of cooperative RRM. An optimised approach to monitoring for getting measurements based on computational intelligence is also proposed. Location information as another advanced methodology for obtaining of correct measurements is also proposed and investigated. Section 2.3 concludes the Chapter.

2.1 Scenarios for Intra-System and Inter-System Cooperation

The general scenario for the investigated cooperative RRM was shown in Chapter 1, Figure 1-1 and Figure 1-2 from two perspectives: inter-system and intra-system cooperation. To ensure simplicity and scalability of the RRM framework, it is proposed here that the RRM functions maybe shared between the two cases. This was shown in Figure 1-4.

Here, it is proposed that RRM functions are divided into **cooperative, generic** and **specific** RRM functions. The interdependencies among these three types of RRM mechanisms was shown in Figure 1-5. Figure 2-1 shows here the location of the three categories of RRM functions in the hierarchy of the physical RAN entities.

Cooperative RRM functions are the functions used for inter-system cooperation. These functions reside at the CoopRRM entity and at the GW. Some cooperative RRM functions may reside also at the RRM server. These functions include inter-system handover, congestion, admission and load control and RAN selection, unified QoS control and authorisation, authentication and accounting (AAA) proxy[1].

[1] The unified QoS control and AAA proxy functions are not a part of the technical details of this thesis but are included for completeness.

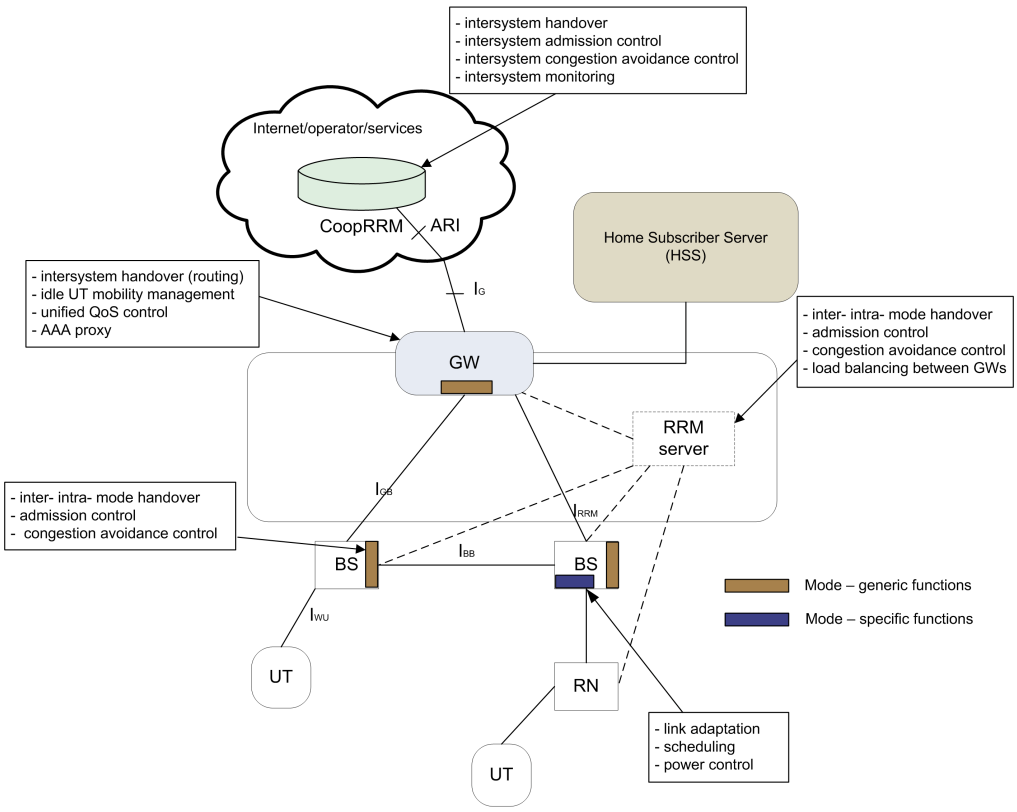


Figure 2-1 Location of RRM functions in the reference architecture.

Generic RRM functions may overlap with some **specific RRM** functions, where the generic RRM will administrate a specific RRM function. An example could be power control where the generic part may specify an interval of admissible power allocations (e.g., regulated by some maximum induced inter-cell interference). The specific part performs the actual power allocation within these limits. Even though some functions may be regarded as being generic, these will still rely on measurements performed by the specific RRM protocols. As the underlying PHY transmission modes may rely on totally different principles (e.g., duplexing schemes) they may perform measurements in a different manner and will have to be reported in such a way that they could be compared to one another. In any case, both the generic and specific RRM functions are intended for intra-system cooperation, and therefore, are proposed to be embedded into the RAN architecture [1]-[4].

Generic RRM functions are shared between the different types of BSs proposed for the reference RAN system (see Figure 1-2 and 1-3) or used for their coordination. Therefore, these functions are also referred to as **mode generic** functions. The generic RRM functions include intra-system handover (both inter- and intra-mode), intra-system congestion, admission and load control, cell/BS selection, flow control and buffer

management. Other functions, not part of the technical proposal of this thesis but included for completeness are spectrum mapping and allocation and link adaptation.

Specific RRM functions are embedded at the RRM server (e.g., resource partitioning) and at BS and RN. These functions include power control, link adaptation and packet scheduling.

The interdependency between the three proposed types of RRM functions was shown in Figure 1-5 as a hierarchical architecture.

The cooperative RRM functions are acting in a centralised approach (i.e., in line with the concept of common radio resource management [5]-[7]) and are highest in the framework hierarchy. The generic and specific RRM functions are acting in a centralised or distributed fashion in accordance with the load of the network (i.e., medium to high loads and low load, correspondingly).

In order to avoid extensive delays in the decision making and signalling overload a solution could be to bring some of the CP functions closer to the BSs, adding more complexity to the BSs and therefore increasing their cost [8]. Assuming that the BS_{WA} overlaps the MA and LA (see Figure 1-2), the extra functionality can be restricted to the BS_{WA}. In [8], the vision for the next generation communication systems is that the cells of the different deployment scenarios will coexist and overlap, either completely or partially. This feature is used in favour of the proposed here RRM architecture as the *mode generic* CP functions that concern the coordination of the different BSs could be moved to the BS_{WA} making it responsible for the control and allocation of resources in the WA cell including all BS_{LA} that fall within its coverage. A requirement for such an approach would be the definition of a communication link between the BS_{WA} and the BS_{LA}. This link could be either wired or wireless. The investigation of the effect of the type of BS communication link is proposed in Chapter 4 in relation to a policy proposed for user context transfer during radio handover and is intended for detailed investigations as a follow up of this research work (see Chapter 7).

Table 2-1 summarises the requirements for the different BSs that serve three deployment areas (BS_{WA}, BS_{MA} and BS_{LA}) in terms of type of physical layer mode, spectrum, mobility and the data rate offered by the deployment cell [3].

The corresponding cooperation mechanism (i.e. mode generic or specific) relies on the type of deployment of the BS, rather than the type of PHY layer mode used by the BS. As shown in Table 2-1, medium- to high-speed users (>70 km/h) can only be served by a WA deployment.

Therefore, the user speed is a parameter that can force UT handover. Future systems may deploy a hybrid information system (HIS) in combination with RRM mechanisms to identify and restore the user profiles, where the mobility parameters such as velocity and mobility as well the relative location with respect to the network deployment are stored [4]. This is also a main capability for use of location-based handover as will be explain in this Chapter 2.

Table 2-1

Types of Base Station	PHY layer mode used	Spectrum	Mobility support	Data rate	Cell size
BS_{WA}	FDD	Licensed	High <350 km/h	Medium (FDD)	High
BS_{MA}	FDD and TDD	Licensed	Medium <70 km/h	Medium (FDD) or Highest (TDD)	Medium
BS_{LA}	TDD	Licensed and unlicensed	Low < 5km/h	Highest (TDD)	Low

It can be assumed that high data rate services of the type of high quality video streaming and highly interactive multimedia, identified as key for future systems [9], [10], would be served only by LA deployment, but an important factor to be taken into account is that the maximum data rate offered to one user depends on the cell congestion, and, therefore, sometimes the maximum data rate could be offered by a MA or WA deployment.

For a system of ubiquitous coverage, intra-system cooperation then would mean cooperation between the BS_{WA}, $BS_{MA,}$ and BS_{LA} and specific to the intra-system cooperation RRM mechanisms would be required.

2.2 Monitoring and Actuation of Cooperation Mechanisms

2.2.1 Measurements

The quality indication for the choice of the most suitable RAN/cell for a given service requested by a user is obtained from measurements on the current and target networks/cells. RRM mechanisms require as much input information as possible. Some examples of the useful metrics are cell load, amount of free capacity, location, velocity and environment of the user, the terminal capabilities, the handover statistics, and so forth. Measurements can be performed either in the BS, or in the UT. A summary of the measurements that are performed in currently available wireless systems, GSM/GPRS/EDGE, UMTS and WLAN, as well as some additional measurements that could be used in inter-system RRM can be found in [1], [11]. Specifically, cell load and free capacity have been defined for UMTS, GSM/GPRS/EDGE and WLAN. Several

important requirements for measurements in the referenced here RAN were deduced from a study on legacy RAT measurements.

For inter-system and inter-mode RRM at least the following information should be obtained from measurements:

- Signal strength measurements;
- Transmitted power measurements;
- Quality measurements;
- Cell load measurements.

Further, each cell of the RAN mode should transmit a beacon that the UTs attached to other systems can listen to and measure, in order to prepare the handover to/from the system or between the different cells [12], [13]. It is proposed here that the inter-system handover should be consistent with the inter-system handover already defined in legacy systems (for instance, handover from UMTS to another RAN) [6]. The UTs connected to one RAN should be able to measure the other RANs efficiently [12].

2.2.1.1 Neighbouring Cell Lists

In order to ease and improve inter-system and inter-mode measurements for RRM algorithms by use of prior knowledge of the relevant parameters, it was proposed in [13] to use neighbouring cells lists. Neighbouring cell lists could be either broadcast, on the RAN broadcast control channel (BCCH), or sent in dedicated signalling messages. For a broadcast neighbouring cell list, all the UTs served by a cell will measure the same neighbouring cells. This can be used if neighbouring cell lists are of low size, (e.g., if the target cells cover the whole primary cell area.) However, if the primary cell covers too many candidate target cells, then dedicated messages are needed, for each UT, to indicate specific neighbouring cell lists corresponding to its position. The detailed advantages and disadvantages of each method are listed in [13] and an elaborate study is available in [14]. Here only the main points are given.

Broadcast of neighbouring cell lists at cell level decreases signalling and can be configured during the deployment process of the network, directly in the operational maintenance controller (OMC) [12], [14]. However, it leads to a quite large definition of neighbouring cells, as the neighbouring cells that are broadcast should cover the whole serving area. Dedicated signalling messages require a controller to dynamically update neighbouring cell lists, UT per UT. It may consequently be quite costly in terms of processing and signalling. Nevertheless, this solution enables to define neighbouring

cell lists dynamically and consequently to take into account several types of dynamic information (e.g., user's service, cells load, and so forth) and is assumed here. The more information is known on the UT and the multi-system environment, the smaller the number of neighbouring cells to measure can be.

During inter-system handover it is required to modify the corresponding RATs neighbouring cell lists definitions when they exist to include their cells. For all measurements cases related to inter-system and intra-system handover, the average neighbouring cell lists size necessary for the broadcast at cell level were evaluated [11], [13], [14], in order to conclude whether dedicated signalling messages are required or not. It appears that dedicated messages are necessary for handover from WA cells/macro-cells to LA cells/micro-cells. The CoopRRM entity then is responsible for computing neighbouring cell lists for inter-system handover, whereas the SRRM entities (see Figure 1-3) are responsible for computing neighbouring cell lists for the inter-mode handover. Location information could be an input to build neighbouring cell lists at the UT level, when dedicated neighbouring cell lists are required [15], [16]. Besides, the size of the neighbouring cell lists could be adapted depending on the current situation; the size can be reduced if the UT is in an emergency situation and needs to perform inter-system or inter-mode handover as soon as possible, or increased (or kept at broadcast level) if the UT has the capability to perform inter-system or inter-mode measurements without degradation.

2.2.1.1.1 Use of Neighbouring Cell Lists for Inter-System Handover

As an example of the use of neighbouring cell lists during inter-system handover, Figure 2-2 shows the proposed general scheme for the algorithm.

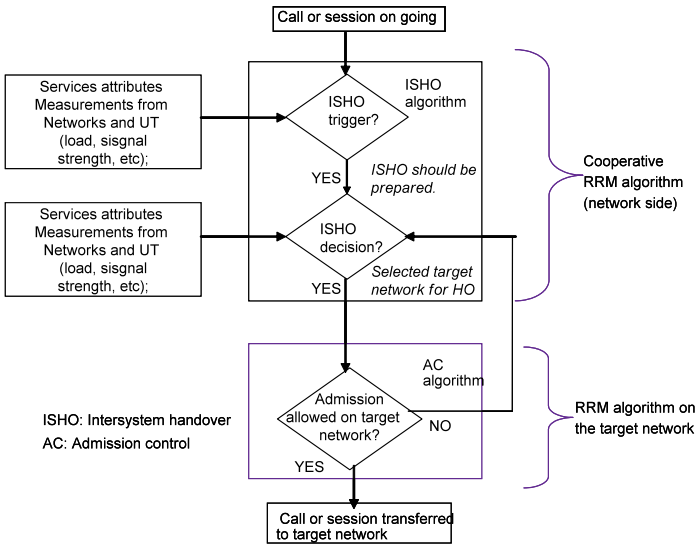


Figure 2-2 General algorithm for inter-system handover.

The proposed inter-system handover algorithm is a cooperative RRM algorithm with two phases: first the decision to trigger the handover and prepare it, then the selection of the most suitable target network to execute the handover to. Both criteria can be based on various inputs from measurements coming from UTs or the networks, or services attributes. Once the target network for handover has been selected by the handover algorithm, the admission control on that target network will check if the call or session can actually be accepted. If not, another system must be chosen, but the handover algorithm should be defined so as to minimize the rejection of calls/ sessions by the admission control.

The information available at each relevant entity (e.g., BS in the reference architecture or RNC in UMTS) can be obtained by the CoopRRM entity via the SRRM entity. Consequently, we assume that the CoopRRM will easily obtain load information on the different cells of the RANs. We also assume that QoS information is available at the CoopRRM for each user (i.e., QoS requirements) and for each system (i.e., the ability to fulfil the QoS requirements of the user). This information is dynamically updated through the SRRM information exchange.

As a consequence, the only limiting factor for the inter-system handover algortihm is the information that can only be obtained through measurements at the UT. *Signal strength* and *signal quality* measurements may be difficult and long to perform, especially in multi-system environment, however, these are required in most handover algorithms. It is necessary to assess that, if the UT performs a handover to a given cell, its signal strength or quality will be enough to ensure the viability of its QoS. In specific cases, such as indoor cells, or during a handover from a WA cell to a LA cell, *signal strength* or *quality measurements* are absolutely necessary in order to choose the most suitable cell for handover. It is not possible to rely on localization information only. Neighbouring cell lists, on the other hand, enable to restrict the number of cells to measure, while still being certain that the *signal strength* or *quality on the target cell* for handover will be enough. For this reason, neighbouring cell lists are integrated into inter-system and intra-system handover schemes, in order to optimize the measurements performed for the handover algorithm. If the handover algorithm requires *signal strength* or *quality information* obtained by the UT, and if the handover trigger is based only on information of the current system, once the handover trigger has happened, the following procedure shall take place in relation to the algorithm proposed in Figure 2-2:

- **Obtain the neighbouring cell list of the UT, on the target system.** We assume that there is only one target system (which may have been chosen previously).
- If the handover algorithm is based on signal measurements only, perform signal measurements on the neighbouring cells of the target system (and also on the primary system if required) and deduce a handover decision.
- If the handover algorithm is based on signal measurements and other information, which may be cell load or QoS, check the other information first. Cell load and QoS information are directly available through the SRRM. If, among the neighbouring cells, some of them do not fulfil the load or QoS requirements of the algorithm, then these cells shall be discarded from the neighbouring cell list. It is not necessary to perform signal measurements on these cells, as the handover algorithm will not choose them for target cell in the end. Consequently, we obtain a sub-list of neighbouring cells on which signal measurements are performed. Then, depending on the algorithm, the target cell is chosen from all the obtained information.

2.2.1.1.2 Optimisation of Inter-System Handover based on Neighboruing Cell Lists

Let assume that after the delay necessary to perform *signal* or *quality measurements*, the load and QoS information on the neighbouring cells has not changed sufficiently and that the load and QoS conditions are still fulfilled.

The measurements procedures with the limiting assumptions would trigger the handover algorithm as shown in Figure 2-3.

If the inter-system handover algorithm is triggered by events on the current and the target systems (periodic triggers), it is necessary to obtain information on the target system periodically. If this information contains signal measurement information, then the following actions must be performed:

- At each period *T,* the neighbouring cell list is updated (if it is necessary);
- If triggering is based on load and/or QoS, check these requirements on the neighbouring cell lists, in order to discard those that do not fulfill them;
- Then perform signal measurements on the remaining neighbouring cell lists.

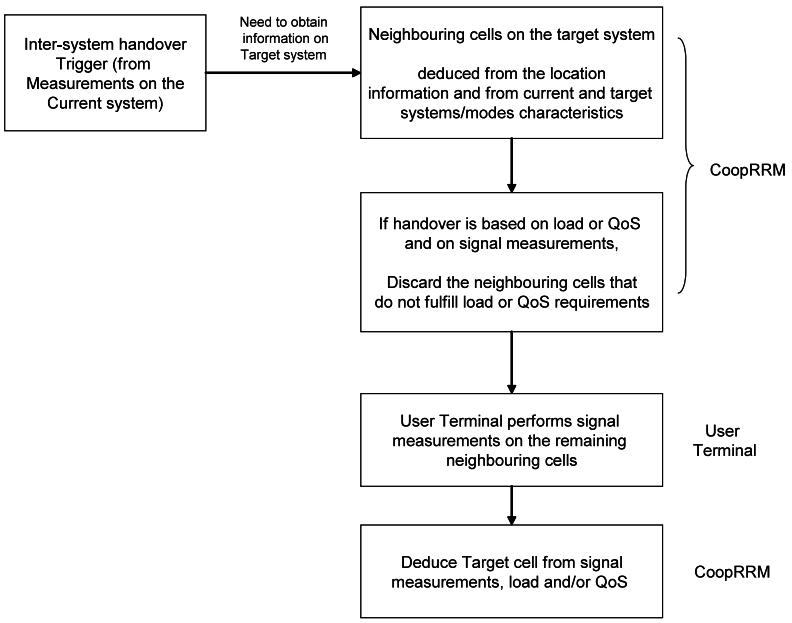


Figure 2-3 Inter-system handover with measurements limitation.

The proposed optimization enables to make use of the CoopRRM functionality efficiently. It is independent of the handover algorithm, and of the final handover decision. The inter-system handover will then be initiated as shown in Figure 2-4. The procedure is also valid for the initialisation of intra-system handover.

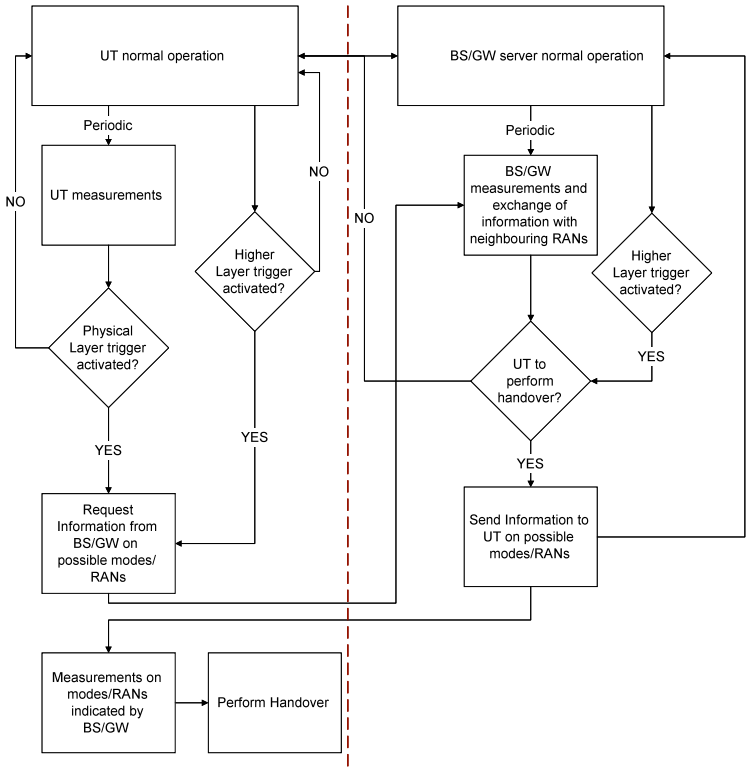


Figure 2-4 Initiation of inter- and intra-system handover with optimized measurements procedure.

Here, the process is described for the case of inter-system handover between the reference IMT-A system and UMTS as the legacy system. However, all three types of RRM algorithms are involved in the process. The following information exchange is required for establishing the new connection on the UMTS system:

- The mode-specific RRM residing in the BS detects a situation in which the UT service requirements cannot be satisfied in any other transmission mode (e.g., loss of coverage or QoS degradation).
- The BS sends an indication message with measurements and statistics to the GW/CoopRRM along with an inter-system handover request.
- The CoopRRM uses periodic or ad-hoc measurements on the other (legacy) RAN candidates to take a decision on handover. In case of ad-hoc measurements, the CoopRRM sends a measurement request to the specific RRM entity ($SRRM_L$) of the target legacy RAN, and this entity answers with a measurement report.
- If the CoopRRM decides to handover to UMTS, the CoopRRM sends a *HO_request* message to the UMTS $SRRM_L$, indicating the identity of the UT and other useful information (e.g., QoS requirements).
- $SRRM_L$ sends a *Hard_HO_request* message to L3 of the UMTS core network.

L3 of the core network sends a *Hard_HO request* message to L3 of the UTRAN that starts the transition from idle mode to *CELL_DCH* state for that UT.

Consequently, a new message between the CoopRRM and L3 of the CN (or of the UTRAN) is exchanged, as well as a new message between the CoopRRM and the $SRRM_L$. This is shown in Figure 2-6. Another approach would be for the CoopRRM to only suggest candidate RANs/cells for inter-system handover to the inter-system handover functionality residing in the BS/GW[2]. Then, the final decision is taken by the BS/GW where all information about the candidate cells and candidate RANs must be collected before making a decision.

The message exchange for the measurements during inter-system handover is shown in Figure 2-5. Each system (i.e. RNC in the case of UMTS) will have to send periodically or on request a message containing measurements information to the SRRM, which will then have to send another message with this information to the CoopRRM.

[2] This was explained partially in Chapter 1

The efficiency and the delay will depend on the location of the SRRM. The closer it is to the RAN, the less the delay from signalling and the faster the decisions. If information to the SRRM is sent periodically without any filtering, then the SRRM only sends information to the CoopRRM on request.

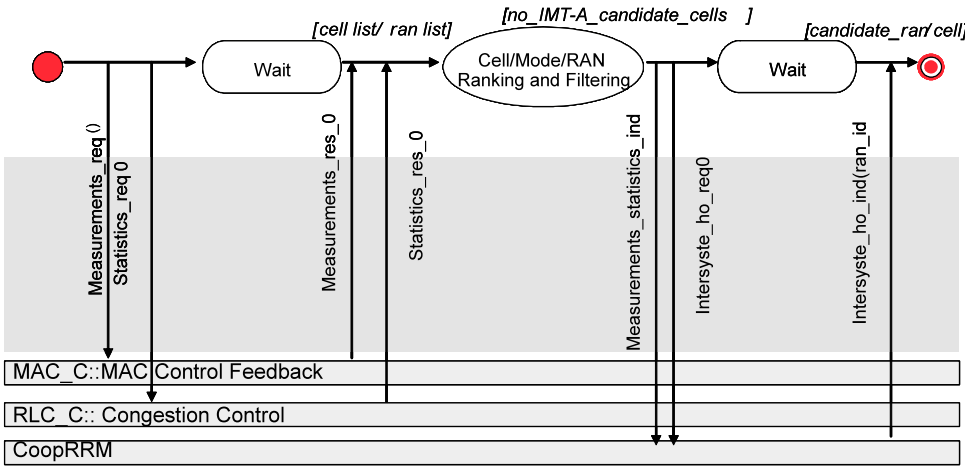


Figure 2-5 Message exchange during inter-system handover.

This is how the functions for inter-system cooperation can be brought closer to the radio interface. This would decrease the signalling load between the SRRM and the CoopRRM.

2.2.1.2 Use of Measurements Strategy for the Initiation of Handover

In the following, three algorithms are proposed based on the adopted measurements strategy in support of inter-system handover. The handover is triggered based on measurements on the current system only. The measurements are used to establish the coverage criteria status and the load status. It is not necessary to perform measurements on candidate target system(s) before the handover decision has been taken.

2.2.1.2.1 Handover from Current System/Cell to Target System/Cell Based on Coverage Criteria

IF $M_{Best,Current} < Th_{Current,1}$, where M is signal strength and Th threshold for coverage;

THEN measurements are triggered on the neighbouring cells identified for the target system.

IF $M_{Best,Current} < Th_{Current,2}$ and $M_{Best,Target} > Th_{Target}$;

THEN handover is performed to the best measured neighbouring cell.

With this algorithm, as only coverage information is used, it is not possible to restrict the neighbouring cell lists because of the need for load or QoS information. Measurements on all neighbouring cell lists will be performed. It is consequently very important, for this handover algorithm, to define neighbouring cell lists as accurately as possible.

2.2.1.2.2 Handover from Current System/Cell to Target System/Cell based on Coverage and Load Criteria

Two possible algorithms are proposed. The first one does not use load as discriminating information, but as ordering information. Consequently, it is not possible to use it in order to discard neighbouring cells.

AlgoLoad(1)

IF $M_{Best,Current} < Th_{Current,1}$

THEN ask for the list of neighbouring cells on the target system and perform measurements on these cells.

If $M_{Best,Current} < Th_{Current,2}$ and $M_{celli,Target} > Th_{Target}$

Then

Handover will be performed to $Cell_i$ of the target system with the lowest load among the cells that have sufficient coverage level (defined by Th_{Target}).

The second algorithm uses load (i.e., L) as a criteria to only keep cells with low enough load. **AlgoLoad(1)** can be rewritten in order to decrease the number of neighbouring cells to measure.

AlgoLoad(2)

If $M_{Best,Current} < Th_{Current,1}$

Then ask for the list of neighbouring cells on the target system.

If $M_{Best,Current} < Th_{Current,2}$

For each of these cells $Cell_i$, if $L_{Celli,\ Target} > Th_{load}$

*Then discard $Cell_i$ from the list of neighbouring cells.

*Perform measurements on the remaining neighbouring cells.
*Then keep $Cell_j$,

if $M_{cellj,Target} > Th_{Target}$

*Perform handover to $Cell_j$ with the best level.

Because handover triggers have to be signalled to the UT, it could be a benefit to be informed prior to actually reaching the cell. Outside the cell coverage area no measurements would be available.

It is very important to accurately use the measurement information in relation to handover. If a trigger is issued too early, the number of unnecessary handovers will be higher and this is cost-inefficient. Therefore, it is necessary to identify correctly the threshold value, which would be used as a reference value for comparison of the results from the performed measurements. The proposed above algorithms have been used for different threshold values to show the claimed requirements. The results are shown in terms of distribution of the handover triggers in Figure 2-6.

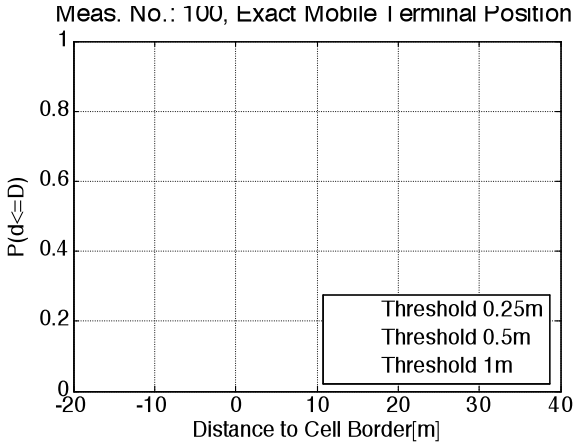


Figure 2-6 Distribution of handover triggers for different threshold values.

It can be seen that increase in the threshold increases the number of handovers that are triggered too early. With low thresholds it is possible that triggers are generated very late. This due to the fact that within the cell the measurement density is not constant, therefore some jitter is experienced. If the threshold drops down and is within the same order as the jitter it is possible that the threshold is passed very late.

Because handover triggers have to be signalled to the UT, it could be beneficial to be informed prior to actually reaching the RAN/cell. Outside the cell coverage area no measurements would be available.

2.2.1.3 Key Performance Indicators (KPIs)

The KPIs are a common way to provide knowledge about the network or link status. The KPIs include relevant to both radio and network performance information. This information is normally obtained by performing periodic measurements. KPIs are a set of measurements used to keep track of a network status over the time. T Therefore, KPIs have been used in relation to the proposed here RRM algorithms.

The KPIs are composed of several raw counters or other measurements collected from the network because a single measurement can be too detailed to be used as a KPI. KPIs are split in two types depending on whether they describe the network's resources or the delivered QoS. The main KPIs related to QoS can be measured in any type of packet-switched network.

From a mathematical point of view, a KPI is a function F: $R_n > R$ such that [17]:

$$KPI = F(reward_1, \ldots, reward_n) \quad \textbf{(2-1),}$$

where $reward_i$ is a performance variable. A performance variable is a generic definition that can be used to represent dependability and performability variables as well. It is strictly related to the modeling tool, in which it is calculated. A performance variable allows for the specification of a measure on one or both of the following:

- The states of the model, giving a *rate reward* performance variable. A rate reward is a function of the state of the system at an instant of time.
- Action completions, giving an *impulse reward* performance variable. An impulse reward is a function of the state of the system and the identity of an action that completes, and is evaluated when a particular action completes.

A performance variable can be measured at an instant of time, measured in steady state, accumulated over a period of time, or accumulated over a time-averaged period of time. Once the rate and impulse rewards are defined, the desired statistics on the measure must be specified (*mean, variance, distribution* of the measure, or the *probability* of the measure falling within a specified range) [17].

The most important KPIs used for support of the proposed here RRM framework are the *delay*, expressed as the time needed for one packet of data (or a flow) to get from one point to another; the *jitter*, expressed as the delay variation of the received packets (inter-RAN flows) over time; the *peak user data throughput,* expressed as the maximum rate achieved during the transmission of data in the network; and the *mean user data throughput,* expressed as the average rate achieved during the transmission of data in the network [16], [18]. Here, one simple example of how to model the delay for the purpose of assessing its dependence on the network load is presented [19].

As the delay varies exponentially with the load of the network [20] a dependency can be derived to obtain a relation between load and delay. In a low network load situation, the delay value (τ) can be represented as a typical delay (τ_{typ}). When the load increases and gets in the congestion zone, the delay value then augments very quickly.

The formula considers the influence of a *congestion threshold* parameter (*CT*) that shows when the congestion zone will be reached. Once this critical value has been reached, the CoopRRM entity will receive a request for handling the arisen congestion situation and an algorithm will be activated. Assuming that no significant change in delay may occur before the 40% load value is reached and that the higher delay value (i.e., for the 100% load value) must remain coherent for the chosen scenario, the delay can be expressed by the following Equation (2-2):

$$\tau = \tau_{\text{typ}} + 120.\text{e}^{\frac{L-\gamma}{12}} \quad \textbf{(2-2);}$$

where τ_{typ} is the typical delay value in ms; L is the load value in percentage of the total capacity of the cell; γ is a parameter, depending on the chosen *congestion threshold* and is expressed in percentage of the total capacity as given by Equation (2-3)

$$\gamma = 40 + \frac{CT}{2} \quad \textbf{(2-3);}$$

The congestion threshold, *CT*, is the load value, expressed in percentage of the total capacity, chosen to identify a congestion situation, and is used to indicate the upper congestion limit. The load, *L,* is defined in a generic way, as a function of the total capacity and is given by:

$$L_n = \frac{\sum_{i=1}^{N_{nu}} DR_i}{C_n} \quad \textbf{(2-4);}$$

where L_n is the load of the nth cell; C_n is the total capacity of the nth cell; N_{nu} is the total number of users running applications in the nth cell; and DR_i is the data rate of the ith user.

The delay variations in a network with $\tau_{typ} = 20$ ms depending on the network load, *L,* are shown in Figure 2-7.

For congestion thresholds defined to be very high (e.g., 85%), the delay variations are less obvious because the network will be able to handle higher loads. For low congestion thresholds, the delay will be more obvious, and this would trigger an RRM algorithm request sooner. Therefore, for higher accuracy, it is very important to define the initial thresholds with a good prior knowledge of the network behaviour in different scenarios.

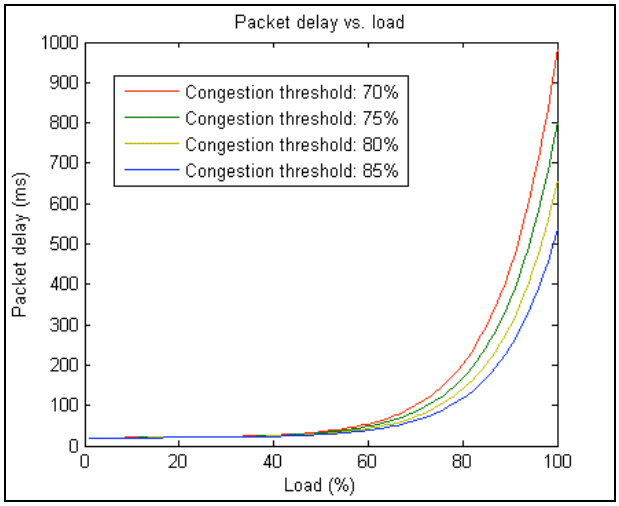


Figure 2-7 Variation of the delay caused by different loads on the network.

2.2.1.3.1 KPI Calculation for Cooperative RRM

The main KPIs related to cooperative RRM can be measured in any type of packet-switched network. The KPIs calculated for the performance of the reference IMT-Advanced RAN serve as an indicator for the user-perceived QoS and achievable system capacity. The KPIs are are defined as follows:

- **Delay [ms].**

The delay, (which can be referred to as latency in the case of WLAN or a system operating in a short-range mode), expresses the time needed for one packet of data to get from one designed point to another. The round-trip delay is measured by the time taken for sending a packet that is returned to the sender. From this, the one-way delay can be calculated, being half of the round-trip delay. A delay much longer than expected indicates congestion in the network. The way the delay is modelled here was described in Equation 2-2.

- **Jitter [ms]**

Jitter represents the delay variation of the received packets over time. Packets sent at a constant rate are not necessarily received at a constant rate, due to the network behaviour (e.g., a congestion situation). Jitter is the measure in time of the irregularity of the packets transmission. Jitter effects can be cancelled through the use of a buffer at the receiving end.

Several formulas for jitter calculation can be defined. Jitter can first be calculated as a raw spreading of the delay around the expected delay:

$$Jitter = \frac{1}{N}\sum_{n=1}^{N}(\tau_n - \tau_0) \quad \textbf{(2-5);}$$

where N is the number of transmitted packets; τ_n is the delay in seconds of the nth received packet; and τ_0 is the expected delay in seconds.

If no expected delay is available, another reference must be chosen, for example the delay of the first received packet. The formula becomes:

$$Jitter = \frac{1}{N-1}\sum_{n=2}^{N}(\tau_n - \tau_1) \quad \textbf{(2-6);}$$

where N is the number of transmitted packets; τ_n is the delay in seconds of the nth received packet; and τ_1 is the delay in seconds of the first received packet.

Jitter can also be calculated with reference to a *mean delay* of the previously received packet, using a recursive formula:

$$\forall n \in [\![1, N]\!], Jitter(n) = \tau_n - \frac{1}{n-1}\sum_{i=1}^{n-1}\tau_i \quad \textbf{(2-7);}$$

where N is the number of transmitted packets; τ_n is the delay in seconds of the nth received packet; τ_i is the delay in seconds of the ith received packet.

Equation 2-7 has the advantage of providing a value of jitter for each received packet, without having to wait for the end of the transmission of the group of packets. This can be interesting especially because the information should be available at each moment of time. Moreover, it may be important to calculate the *absolute jitter*. Indeed, with the formulas defining the raw jitter (see Equations 2-5 and 2-6) the delays of packets arriving late can be minimized by the delays of packets arriving early, and information might be hidden. That is why the absolute jitter is also defined, which will have in any case a value equal or greater than the raw jitter. The formulas are given, respectively, for jitter with reference to an expected delay (see Equation 2-8); jitter with reference to the first arriving packet delay (see Equation 2-9) and jitter with reference to mean delay (see Equation 2-10):

$$Jitter = \frac{1}{N}\sum_{n=1}^{N}|\tau_n - \tau_0| \quad \textbf{(2-8);}$$

$$Jitter = \frac{1}{N-1}\sum_{n=2}^{N}\left|\tau_n - \tau_1\right| \quad \textbf{(2-9);}$$

$$\forall n \in [\![1, N]\!], Jitter(n) = \left|\tau_n - \frac{1}{n-1}\sum_{i=1}^{n-1}\tau_i\right| \quad \textbf{(2-10).}$$

- **Peak user data throughput [bps]**

Peak user data throughput is the measure of the maximum rate achieved during the transmission of data in the network. This KPI must refer to a single user. For this measurement, an *instantaneous user data throughput* must be available from the network, i.e., an instantaneous value of the data rate for each user.

The *peak user data throughput* calculation is based on the dependency $PUDT = \max(IUDT(t))$ where *IUDT*(t) is the *instantaneous user data throughput function,* in bps.

- **Mean user data throughput [bps]**

Mean user data throughput is the measure of the average rate achieved during the transmission of data in the network. This KPI must also refer to a single user.

The calculation is made by comparing the size of the transmitted data with the time of transmission of these data, both for uplink and downlink. The calculation may also be done with an integration of the *instantaneous user data throughput* function. The *mean user data throughput* is calculated separately for the UL and DL (see Equation 2-12 and Equation 2-13, respectively):

$$MUDT_{UL} = \frac{uploaded\ data\ payload}{upload\ time\ for\ data\ transf\ e} \quad \textbf{(2-11);}$$

$$MUDT_{DL} = \frac{downloaded\ data\ payload}{download\ time\ for\ data\ transf\ e} \quad \textbf{(2-12);}$$

Interesting KPIs concerning the network status are the *available bandwidth* and the *throughput*. In the context of the RRM framework proposed here, the bandwidth does not represent a range of frequencies but is employed in a data rate sense. A given frequency range can bear a corresponding data rate, depending on the used coding scheme and multiplexing technique, however, *bandwidth,* in the data rate sense, is the speed at which a network element can forward traffic. It has two characteristics –

physical and available, and both of them are independent of end hosts and protocol types.

The *physical bandwidth* or *capacity (C)* is the maximum number of bits per second that a network element can transfer. The physical bandwidth of an end-to-end path is determined by the slowest network element along this path. Here, an *utilisation (U)* factor is defined (see Equation 2-13) indicative of the percentage of capacity consumed by the aggregated traffic on a link or path:

$$U = \frac{Traffic}{C} \quad \textbf{(2-13);}$$

The *available bandwidth (A)* is the capacity minus *U* as defined in Equation 2-14 over a given time interval:

$$A(t_s, t_e) = Capacity - Traffic$$

$$\Leftrightarrow A(t_s, t_e) = C \times (1 - U) \quad \textbf{(2-14);}$$

Where t_s is the time at which the measurement starts and t_e is the time at which the measurement ends.

- **Throughput [bps].**

Throughput is the amount of data that is successfully sent from one host to another via a network. It may be limited by every component along the path from source to destination host, including all hardware and software. Throughput also has two characteristics – *achievable* throughput and *maximum* throughput. *Achievable throughput* is the throughput between two end points under a completely defined set of conditions, such as transmission protocol, end host hardware, operating system, tuning method and parameters, etc. This characteristic represents the performance that an application in this specific setting might achieve. Therefore, the available bandwidth is a measurement that indicates whether there are still resources in the network that the users can exploit. The achievable throughput can be low even if there is still available bandwidth, for example, this is the case when a network element is limitative.

There are two ways of estimating the available bandwidth: the *passive* measurement, which consists in using the existing data transmission history, and the *active probing*, which consists in creating the situation in which it will be possible to measure the available bandwidth. The principle is that the sender sends a pair of packets

echoed back by the receiver. By measuring the changes in the packet spacing, the sender can estimate the bandwidth properties of the path [21]. The available bandwidth is a dynamic property depending on many factors [21].

- **Network load [%].**

The *load* of the network describes how much the network is utilised over time, in term of resources. Existing definitions for the calculation of the load of a network are very different depending on the type of network dealt with [21]. In GPRS, it is only the ratio between available time slots and the total number of time slots. In WLAN, it is done either like in GPRS or with a ratio between collisions and transmissions. In UMTS, the definition is much more complicated as it embeds the influence of noise and of interferences, with different calculations for UL and DL.

In order to simplify and generalise the load computation that would allow for assessment of the proposed RRM framework in generic terms, the load is defined as in Equation 2-4, i.e. only with respect to the bandwidth metric in the data rate sense. In this way, a simple definition is obtained that can be used with all modes. Load then describes the utilisation of the network capacity, in other words, the quantity of occupied bandwidth.

With this assumption, the value of the network load is obtained by adding the value of the bandwidth used by each user at a time.

KPIs are used for the monitoring process for the implementation of the proposed RRM framework as a real-time simulation platform. The calculation procedure for the KPI aggregation is proposed in Chapter 5.

2.2.1.4 Triggers for Cooperation Mechanisms

A trigger is a network measurement that indicates changes of setup or surrounding conditions based on which a cooperation mechanism is activated, if the pre-determined threshold has been reached. There are different types of triggers, physical-layer based L2 triggers, algorithm-based L2 triggers, and so forth. A detailed analysis of the selection procedure of triggers is available in [13]. Here, triggers are proposed for the proposed cooperative RRM framework.

2.2.1.4.1 General Triggers

The triggers used in support of the proposed RRM framework have been classified into two main groups. The first group consists of triggers that necessitate handover and therefore if a handover does not take place the call will be dropped. The second group contains triggers that can cause a handover but if this handover is not performed the call

will not necessarily be dropped. The two trigger groups are summarised in Table 2-2 for the case of inter-and intra-system handover.

Table 2-2 Groups of Triggers for Inter-system/Intra-system Handover

Triggers that necessitate handover	Triggers that cause a handover
Signal strength	Cell load of current and target networks
Interference level	Compatibility of user preferences
BER/PER	Data rate requirements
Carrier-to-interference(CI) ratio	QoS requirements and violations of these
UT velocity	Policy of the operators
	Location of UT

2.2.1.3.2 Triggers in an Interference Constrained Environment

In an interference-constrained environment, assuming cognitive-radio enabled UTs, several layer 1 and layer 2 performance measures can be used to trigger inter-system or intra-cell handover (i.e., between BSs serving different deployment areas).

The signal level measure alone may not necessarily reflect the level of interference between the devices operating in the same band and the UT that needs to handover, therefore, the number of the packets retransmitted at the MAC layer can be used as a measure of the number of packets that have been discarded due to packet collisions. When the number of packets exceeds the threshold value, the UT may trigger a handover.

This measure captures the percentage of the packet loss at the receiver. In addition, this measure can indirectly provide information about the packet collisions at the receiver at the other end. For example, if a mobile device correctly receives data packets but observes that the BS is trying to send the same data packets several times, it means that the acknowledgements are lost at the BS. In that case, the cognitive-radio enabled device can trigger a handover [22]. This decision process can be assisted further by capturing the specific application requirements in terms of bandwidth, delay, and packet loss that also relate to the overall QoS the user expects that we can refer to as quality of information (QoI) [22], [23]. Different threshold requirements can be devised and quickly evaluated by means of inference and learning techniques implemented at the convergence layer and these results can be made available at the lower layers (e.g., link layer). For example, to support better a real-time video streaming application, the delay between each packet received can be monitored and a handover can be triggered if the delay variance (jitter) goes beyond a pre-defined threshold.

2.2.2 Enhanced Collection of Measurements

The monitoring process for the collection of measurements can be provided with *a self-learning* capability that uses outcomes of decisions and observations and learns from these to characterize and predict the system or user behaviour and thus decrease the number of required handovers, i.e., user-context transfers. Intelligent monitoring and self-learning is possible by introducing the following additional functionalities:

- Cognitive sensing in the radio segment;
- Collecting and mining information and status from the network and full detection functions;
- Cognitive learning at decision level (located in the CoopRRM entity).

For the implementation of intelligent monitoring within the RRM framework, (see Figure 2-1) it is proposed to provide the BS and GW nodes with autonomous management functionalities based on use of cognitive radio technology and cognitive routing, elevating the process of traditional network management to a *cognitive state* for improved network performance (e.g., improved handover performance) [22], [24]. The proposed implementation is shown in Figure 2-8.

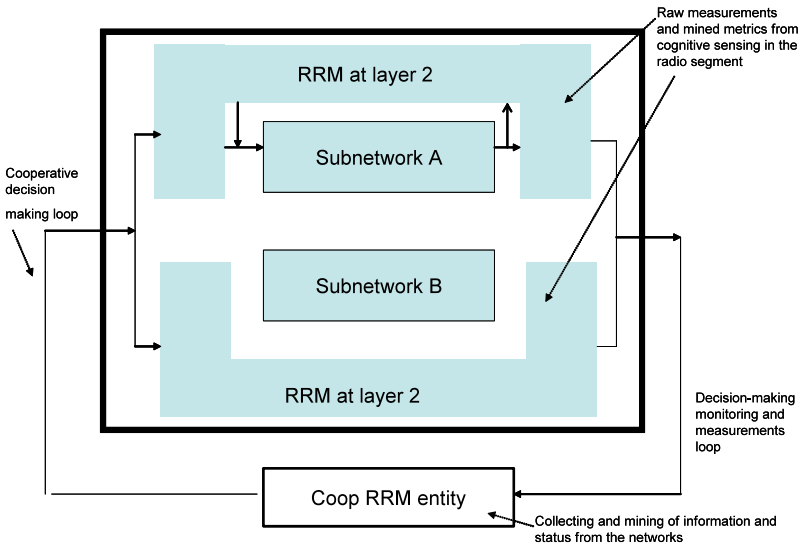


Figure 2-8 Proposed implementation of 'cognitive' monitoring.

Based on the input received from the monitoring sub-network, the main monitoring module and the CoopRRM perform decision-making processes in order to identify suitable strategies to relief the effects of the congestion that can occur when the traffic load is increased with new admission requests. To that, they have available a set of RRM techniques, (RMTs), which represent the means by which the allocation of

resources to the incoming traffic can be arranged in order to optimise resource utilization.

In order to achieve a quick reaction to the network overloads, the approach followed is to have a more refined decision-making within the monitoring module (see Figure 1-4). To reduce the complexity to a manageable level, it is proposed to introduce complementary decision-making functionalities into each module of the RRM engine, including the enhancement of the CoopRRM module.

Figure 2-9 shows the implementation of a learning block at the CoopRRM entity. In order to elevate the network management process to a cognitive state, the decision making process is assisted by use of fuzzy logic rules.

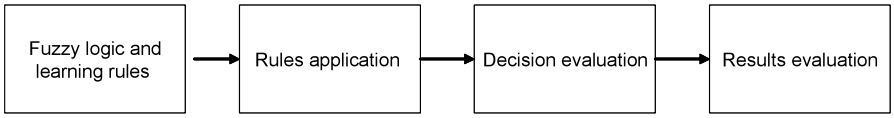


Figure 2-9 Enhanced decision making for the collection of measurements by use of fuzzy logic rules.

Fuzzy logic [25] is a simple and fast solution to provide a conclusion from imprecise, noisy or incomplete inputs. Fuzzy logic is based on simple "IF *X* AND *Y* THEN *Z*" rules rather than complicated mathematical models. System behaviour can be tuned, simply by modifying the appropriate rules and it is possible to compare quantities from heterogeneous RANs. In complex systems, fuzzy models, based on simple IF-THEN rules, give more easily assimilated information than precise models. Nevertheless, rules definition requires a good knowledge of the systems and prior field experience. Therefore, learning techniques are needed to provide full knowledge about the system parameters.

Fuzzy logic has been proposed here to enhance the monitoring process for the aggregation of KPI values.

The proposed implementation of an enhanced CoopRRM block based on use of fuzzy logic is shown in Figure 2-10.

Cooperative RRM techniques will be activated in the case of an inter-system interworking (e.g., inter-system handover). Fuzzy logic rules for cooperative RRM will be defined as an input to the fuzzy logic controller. Two typical fuzzy control systems are prevalent currently, the Mamdani type [25] and the Takagi-Sugeno (T-S) type [25], [26]. Mamdani type fuzzy systems employ fuzzy sets in the consequent part of the rules, whereas the T-S fuzzy control systems employ function of the input fuzzy linguistic variables as the consequent of the rules.

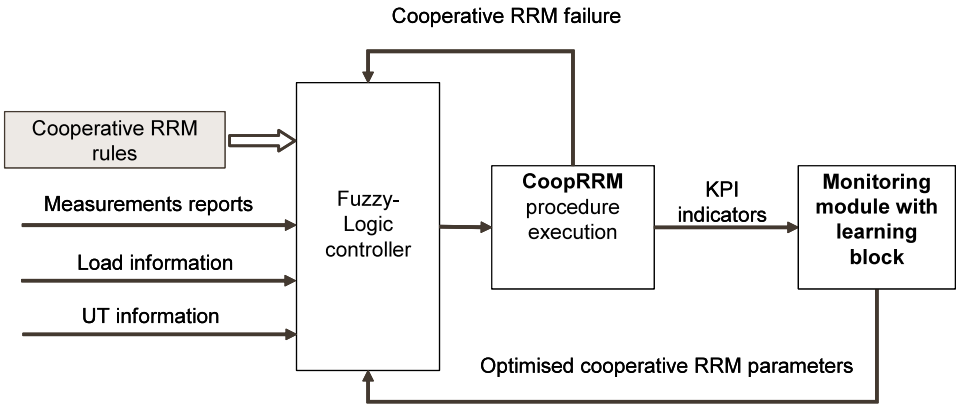


Figure 2-10 Enhanced decision making at CoopRRM entity based on use of fuzzy logic.

Here, a Mamdani type fuzzy control system for cooperative RRM is proposed. This means that the rule format for the cooperative RRM (RRM-g, see Chapter 7) rules will be defined as follows:

RRM-gR_j: **IF** x_1 is $A1^j$ **and** x_2 is $A2^j$ **and** ……,**and** x_n is An^j
THEN y is Bj, j= 0, 1, 2, …., M,

where x_i for I 0 1, 2, …, n are linguistic input variables;
Ai^j *for i=1, 2, ..., n* are input fuzzy sets;
y is a linguistic output variable;
Bj is the output fuzzy set, and
M is the number of fuzzy rules.

The obtained measurements and monitoring information are detailed and membership functions per measurements are determined with values as relative members of a given set of values within an interval referred to as *fuzzy set* [25], [27]. A *fuzzy set* ***A*** is a set of ordered pairs given by the relationship in Equation 2-15:

$$A = \{(x, \mu_A(x)) \quad x \in X\} \qquad \textbf{(2-15);}$$

where X is a universal set of objects and $\mu_A(x)$ lies in the closed interval of [0, 1] and is the grade of membership of the object x in A. Then the membership function $\mu_A(x)$ is characterized by the following mapping as defined in Equation 2-16:

$$\mu_A : x \to [0,1], x \in X \qquad \textbf{(2-16);}$$

where x is a real number describing an object or its attribute and A is the subset of X.

A detailed explanation of how membership functions are constructed and their representation by mathematical functions is available in [25].

When the rules are fed into the Mamdani fuzzy controller, the fuzzy inference engine will compute initially the truth values, then the contribution of each rule, and finally the rules will be aggregated to generate a fuzzy control signal.

The rules are applied and mapped to the logical inputs. The output is a set of triplets made up of a cooperative RRM-g decision and an evaluation degree. Based on the evaluation degree, a decision is taken for the most suitable outcome of the requested cooperative RRM-g process.

As an example of an enhanced decision-making process, a congestion handling procedure is proposed. In this case the decision-making system is used to trigger an algorithm that guarantees QoS to new and already connected users. The following input is decisive for triggering of the algorithm:

- Congestion / No Congestion;
- Throughput (LOW, HIGH);
- History of rejected user (LOW, HIGH);
- Number of occurrence "*decrease lowest priority of undecreased session bit rate by half*";
- Number of occurrence "*Drop lowest priority*".

The resulting actions will be:

- *Reject / Accept* user (from HO or from new session);
- *Increase* OR *Decrease* lowest priority undecreased session bit rate by half;
- Drop lowest priority.

The input parameters can be processed by means of the fuzzy logic rules. Possible rules for congestion detection are shown in Table 2-3.

Table 2-3 Rules for Cooperative RRM for Handling of Congestion

	No Congestion	**Congestion**
Throughput < Capacity threshold AND *Load < Congestion Threshold*	Increase bit rate starting from high priority session/user to the lowest one *	CONGESTION RESOLUTION PROCESS
Throughput = Capacity threshold	N/A	CONGESTION RESOLUTION PROCESS

The following decision-making rule is applied:

***IF** congestion is LOW **AND** system throughput < capacity **AND** load < congestion threshold **THEN** increase bit rate until Throughput = Capacity.*

In case the congestion cannot be handled otherwise then by performing inter-system handover then the decision will be based on the following rule:

***IF** (current system = RAN1) **AND** (UT velocity = LOW) **AND** (RAN2 coverage = MEDIUM OR HIGH) **AND** (RAN2 load = LOW OR MEDIUM) **THEN** (handover to RAN2).*

In this case we assume that the parameters triggering handover are the UT velocity[3], the coverage and the load. This is shown in Table 2-4.

Table 2-4 Rules for Inter-System Handover

Current signal strength	Current load	UT velocity	Best target signal level	Best target load	Handover decision
LOW		HIGH	>LOW	<HIGH	YES
	HIGH		>LOW	<HIGH	YES
HIGH	LOW	HIGH			NO
			LOW	HIGH	NO

The advantage of use of fuzzy logic-based rules is that this approach complies with the generic character of the proposed RRM framework. In addition, other advantages are that the decision response time can be reduced, and the overall TCP throughput can be enhanced [16].

Another way to autonomously perform measurements in the destination network, in order to collect required for handover decisions data, is by the UT itself. If this procedure shall take place during an ongoing connection, two transceivers are required, which enhances the complexity of the UT and is not regarded in further detail here. If no ongoing connection is active, the UT may switch to another network in order to derive respective measurements at arbitrary times. Nevertheless, to prevent the UT from being paged from its current system while scanning another one, respective signalling indicating some kind of sleeping and temporary non-availability is necessary. If the UT demands for up to date information on other networks in order to guarantee the best QoS to the user, the aforementioned signalling/scanning procedure needs to be repeated on a regular basis resulting in transfer overhead that does not even pay off if the conditions in the possible destination network are too bad and thus no handover takes place.

[3] The user velocity in this general case is not critical but needs to be considered when identifying the most suitable target RAN/cell

2.2.3 Measurements based on Location Information

Information about a target RAN/cell can be gathered also by foreign-party-based measurements [2], [4], [28]. The idea is that a nearby located UT of the other system makes a status report and transfers this report by the GW to the currently employed network. Hence, an overview of the conditions of possible destination systems is provided without the need for leaving the current system. Even if the existence of other systems is announced in the broadcast channel, the question remains, which link conditions the UT can expect if it really changes to the announced system. Thus, information about other systems as well as their link conditions needs to be provided. It must be noted that it here it is not explicitly proposed to include inter-system information in current broadcast transmissions. It is proposed to employ measurements taken by other parties. The interesting aspect concerning the gathering of those measurement reports is that they do not need to be rendered explicitly. The idea is to exploit available information, (e.g., signaling information with the original purpose to adjust power control mechanisms or link adaptation). The challenging task is how to process, recycle and supply the information. It was proposed in [2], [13] to use the *Hybrid Information System (HIS)* concept for this.

Figure 2-11 shows the information exchange during a handover based on the use of the proposed RRM framework and employment of the HIS.

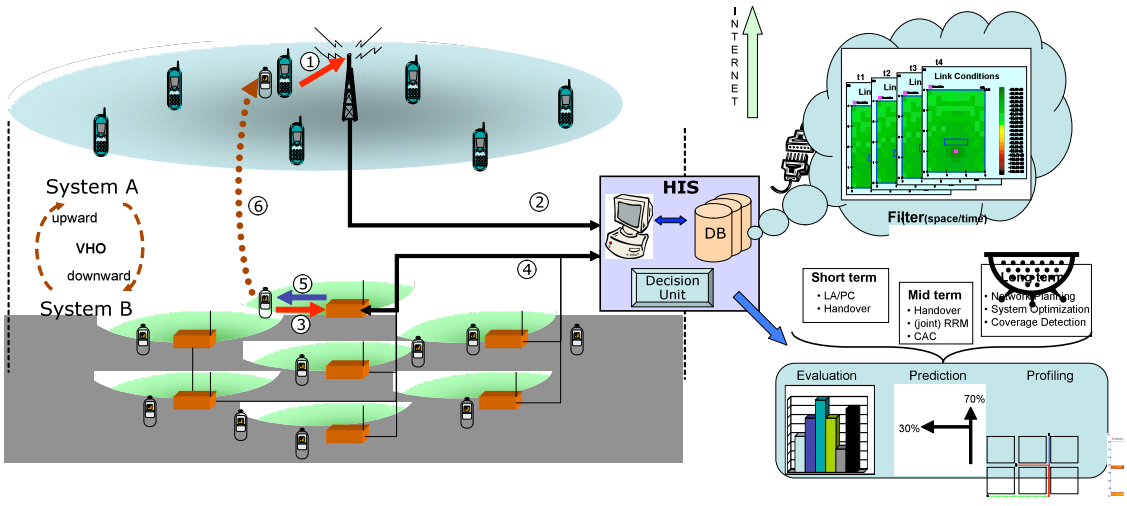


Figure 2-11 Exchange of handover reports between systems by use of location information available through the HIS.

Each active UT reports about the current link condition (1) according to the procedure proposed in Figure 2-4. Together with the measurement report the location of the reporting UT is stored in a database (DB) (2). A UT that intends to perform an inter-

system handover sends a request to its BS, see (3). The BS acquires the corresponding measurement report from the DB, depending on the current location of the UT, (4), and signals the HO decision (respectively related information that allows the UT to take the decision) to the UT (5). The UT can then perform the handover, which is marked by step (6).

Measurements that are inherently available for each system are made available to support the inter-working between the heterogeneous systems. Depending on the new target system and the current location of the UT, the UT is supplied with state reports of the same system type (i.e. intra-system handover) or for the different systems available (i.e., inter-system handover). Handover that exploits the location of the UT for the decision making is referred to as *location-based handover.*

The HIS is both, an intelligent concept facilitating inter-system cooperation and a means to allow for context transfer between different systems. The HIS entails a decision unit that takes into account trigger origins as input and produces handover recommendations (i.e., handover triggers) as output. The advantage is that the HIS is not restricted to local and system specific trigger origins. Besides incorporation of a multiple number of systems, HIS supports load balancing and joint RRM, RRM techniques that are employed in the proposed cooperative RRM framework. Further, due to its backbone connection (see Figure 2-11), specific user preferences may be requested (e.g., from the home network provider) and incorporated in any decision process. Thus, the HIS supports intelligent inter-system-control by combined evaluation of various trigger origins. The elements of the DB proposed for the HIS are described in detail in [1], [2], [13], [14]. Here, it is only sufficient to mention that it allows the storage of *short-, mid-, and long term* data. This is an important property for the proposed here RRM framework because it allows to employ information for decision making based on the specific scenario of the moment. For example, short-term data is employed for real-time requests, while long-term data is employed for less time critical scenarios (i.e., static information for predictable actions).

With this classification, the different sets of data can be applied to different the different types of RRM algorithms using the information from the HIS as input to KPI calculations. To increase the coverage or adapt to different loads in a system, an algorithm that dynamically adjusts the down-tilt of the BS antennae may be employed [29]. This algorithm can use the *mid-term* or *long-term* data as input to its calculations. To support cooperative, generic and specific algorithms (e.g, handover, link adaptation

or power control), *short-term* or *mid-term* data would be used. The classification of the data used for support of cooperative RRM, is given in Table 2-5.

Table 2-5 Types of HIS Data for Cooperative RRM

Short-term data	Mid-term data	Long-term data:
− Support of ongoing handover execution − Decision basis for time critical handover − 'Short' life cycle − Fading, shadowing and other propagation effects included − High variance over time	− based on short-term data − averaged/extracted from short-term data − quasi-static character since short-term fading -> excluded − cell breathing -> still included − shadowing included	− cell breathing -> excluded − based on mid-term − Period: >= 1day − Periodicity given? (e.g., regular football matches) − Used f. network optimization • Coverage • Detection of shadowed areas

It must be mentioned that accuracy is crucial when employing location information for decision making [6], [16]. The less accurate and precise the location information is, the larger the difference between the anticipated, (i.e., retrieved) measurement report, and the real link condition in the target system after the handover.

2.2.3.1 Use of HIS for Cooperative RRM

The HIS information here is used to assist the decision during intra-system handover. Figure 2-12 shows a scenario which requires that an inter-mode handover decision for a user initially connected to the BS_{WA}.

Based on location-related measurement reports the HIS calculates cell borders and signals a UT to handover as soon as the target cell is reached. UTs that are located near the cell borders and that do not move too fast can be identified easily by the HIS. These UTs may in principle be used by other UTs currently outside the cell coverage to enable communication to the other modes (WMs). The HIS informs the UTs that are leaving the cell about a possible topology within their vicinity to establish communication with that attachment point. In that way measurement reports for areas outside the cell coverage may be gathered. These reports may then be used to determine the link quality outside the cell coverage to recommend the modes and hence the possible attachment points to the arriving UTs.

It should be noted that link quality measurements lose much of their relevance with varying position information. To solve this problem in general, each measurement

needs to be associated not only with the position of the UT but also with the current position of different attachment points including fixed and mobile relays (see Figure 2-14).

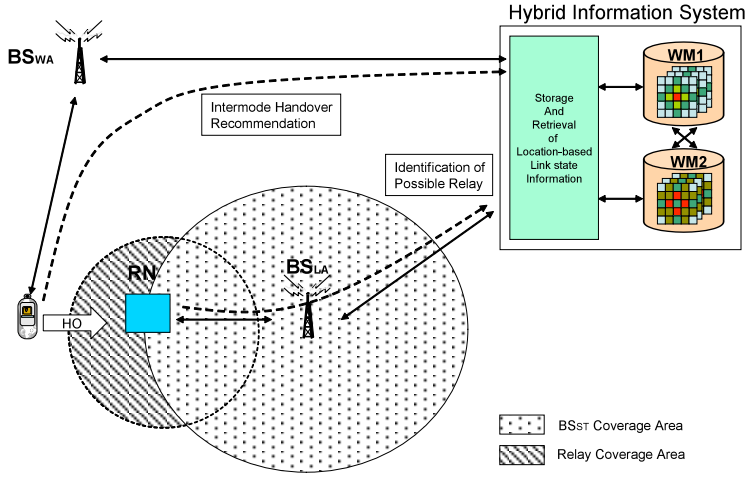


Figure 2-12 Identification of possible attachment point.

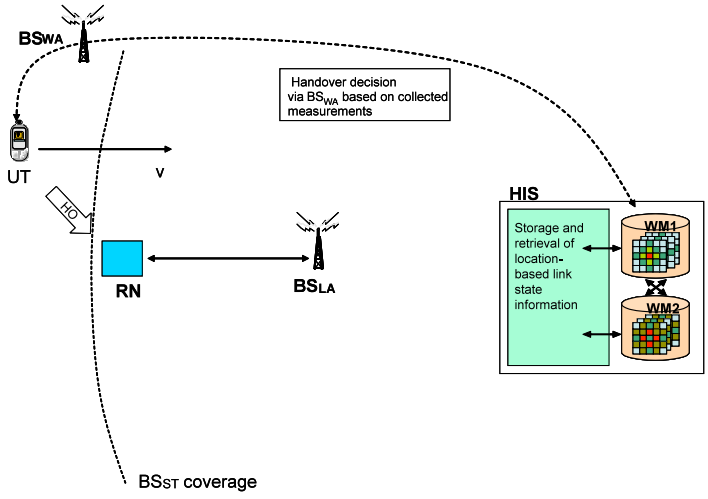


Figure 2-13 Triggering of handover for different deployment scenarios with information from HIS.

To overcome an increase in data complexity, probabilistic statements based on measurements and basic assumptions for the signal range of currently available points can be considered.

The following are examples of how to find the optimal handover points between RNs and BSs. This type of optimisation was not investigated further in the scope of this research work but is mentioned here for completeness:

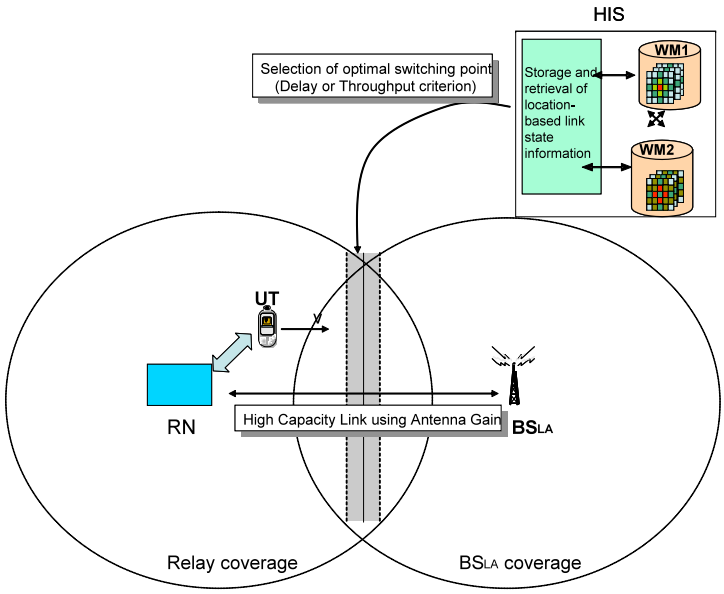


Figure 2-14 Optimization of RN-BS handoff point

- **Frame measurement report.** Frame measurements can be used to easily gather *received power histograms* (RPIs) for different source stations. Thereby measurements are only made during frame transmissions and are each associated with the MAC address of the frame source. This is especially useful to perform measurements for a multiple stations simultaneously. If relay and access point operate on the same channel both signal strengths can be measured simultaneously.
- **Channel load report.** This measurement both takes into account the physical carrier-sense mechanisms (*clear channel assessment – CCA* [36]) as well as the virtual carrier-sense mechanisms (*network allocation vector – NAV* [37]), to determine the current channel utilization. This is especially useful to estimate the available system capacity.
- **Medium sensing time report.** The channel load report gives information on the current channel use. The medium sensing time report gives more detailed information by not only indicating a percentage of the used channel but reporting a histogram of sensing times, which allows for a more sophisticated view on the channel status and thus for an estimation on the expected packet delay [38], [39].

- **STA statistics report.** This measurement can be used to query different counters within the UT, such as retry, multiple retry and failed counters. This allows for information gathering that does not only account for the physical layer but incorporates the link layer as well [40].

It is possible to perform position-based handover decisions using the Centre of Gravity (CoG) algorithm [14], [41]. The CoG algorithm was designed to compensate effects of 'misleading' measurements introduced to the database by erroneous positions. Thereby 'misleading' measurements are measurements that actually have been recorded inside the cell coverage. Due to positioning errors, associated coordinates reported along with the measurements indicate positions outside the actual coverage area. *'Correct'* measurements suffer from the same positioning error but the reported position effectively is inside the cell coverage area. The CoG algorithm exploits the fact that the density of 'misleading' measurements is lower than the density of 'correct' measurements. When a UT is approaching the cell boarder, the algorithm calculates the distance from the terminal to the CoG. The CoG algorithm not only gives a scalar distance, but returns a vector towards the centre of gravity. This allows for estimation whether the UT is moving towards the cell centre or whether it is just passing by. Accordingly, it may be applied in the context of ping-pong handover avoidance [42].

Figure 2-15 shows how the position of the UT can be determined to decide whether the UT is within the coverage area.

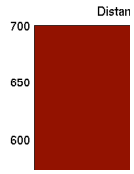


Figure 2-15 Determining the position of the UT by use of CoG.

The decision area is drawn around the UT position. For Figure 2-15 the radius of the area was chosen at 5m. The algorithm is applied to the erroneous positions associated with each measurement report.

Incorporating more information sources into the handover decision will further optimize the resource utilization. This means that more measurements from the attached communication system will be incorporated and that in addition to these estimations of the current network state, the user profiles, the current QoS demands and the operator policies need to be taken into account, too. Systematic combination of both physical measurements and the guidelines that are defined in the operator policies can be successfully handled by a well defined framework. In cellular-network-based positioning the localization process is generally based on measurements in terms of *Time of Arrival (ToA), Time Difference of Arrival (TDoA), Angle of Arrival (AoA),* and/or *Received Signal Strength (RSS),* processed by the network or UT [43].

Another proposed solution is based on the Global Navigation Satellite Systems (GNSSs) [45].

For a general solution a hybrid approach is suitable depending on the UT position and the environmental conditions. Usually, as much as possible, available information sources should be used for positioning. In WA scenarios where good LoS access to the satellites is possible, a GNSS-based solution is the best choice with supporting information and measurements from the RAN. In LA or indoor scenarios, where no satellite signals are available, a pure RAN-based location determination is necessary. In MA, it could be a RAN-based solution where – if available - GNSSs signals are used to improve the positioning of the UT. Again, the limiting factors in these scenarios are determined by non-LoS and multipath effects [3], [13], [14].

Generally, the location estimation can be done within the UT using measurements and information sent by a location service support function, or within the location service support using measurements sent by the UT and/or BSs and relay nodes involved in the location estimation process. The best stand-alone based performance can be obtained by including timing measurements of the UT. In-band timing measurements in cellular-network-based positioning are usually based on TDOA measurements.

Figure 2-16 shows the cumulative density function (CDF) of the estimated positioning errors where the performance of GNSS (i.e., Minn algorithm), investigated in detail in [46], is tested in a cellular network environment under multipath conditions and the parameters given in [44]. The bandwidth is 50MHz, which results in a chip duration of $T_{\text{Chip}} = 20\text{ns}$. From a positioning point of view this chip duration yields an

equivalent chip length of 6m. The system combines GNSS measurements with two TDOA measurements taken from simulations with a MIMO channel model, developed for an IMT-Advanced candidate system [16], in the base coverage MA scenario. All measurements are integrated in an extended Kalman filter (EKF) tracking algorithm where a pedestrian user is assumed [46]. Furthermore, a carrier-frequency offset of $v = 10$ is assumed and is normalized to the subcarrier spacing. NLOS propagation is not investigated here. For positioning the strongest three BSs are used and an averaging is performed over 100 synchronization symbols, which are equivalent to about 0.5s, i.e., every 0.5s new TDOA estimates are available. Additionally, an SNR at the cell edge of -5dB is assumed with a cell radius of R = 500m.

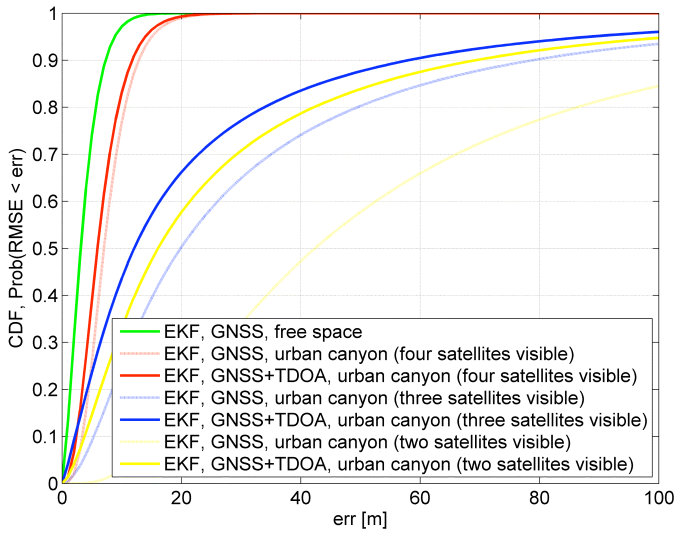


Figure 2-16 CDF for satellite-based positioning, combination of GNSS with two TDOA measurements in MA scenario.

For optimum GNSS conditions (free space), the 90%-error is below 7.5m and it can be seen that the performance gain by additional TDOA measurements is small when enough satellites, (i.e., at least four), are available for LoS. However, for only three or two satellites the performance can be increased and the lack of satellites can be compensated by the RAN TDOA measurements. For instance, in 90% of the cases the error can be reduced from below 80m to below 60m if only three satellites are available, for two visible satellites the performance gain by additional TDOA measurements is even higher.

Except for Cell ID, all possible inputs for the location determination process rely on link-level physical layer measurements. However, all methods provide different performance in terms of accuracy, frequency, reliability, and complexity. Furthermore,

some of them strongly depend on the capabilities of the UT. For instance, AOA requires multiple antennas at the BS and/or the UT, and GNSS requires modifications at least at the UT. If these hardware requirements are fulfilled, the performance of the location determination process will be fulfilled. In [46] it was shown that location information should be provided for all UTs in the RAN.

Use of position information to manage radio resources is attractive but also a sensitive new field. The fact is that exploitation of location information requires new regulative actions, legislation and self-commitments. Supervision to avoid possible illegal and unethical use of personal information is also imperative.

2.3 Conclusions

This chapter proposed scenarios and measurements strategies for the actuation of cooperative RRM mechanisms for inter-system and intra-system interworking. It was shown that cooperation is required on three levels, supported by cooperative, mode generic and mode specific RRM mechanisms. Such an approach has benefits for emerging communication systems, because the proposed framework ensures the coexistence with legacy systems. The proposed RRM framework is based on the proposal for a flat distributed RAN envisioned for next generation systems and already approved for LTE. Further, it ensures that active inter-connections between relevant RRM entities are maintained at the desired cooperation level. The inter-node connectivity provides interworking, which in turn provides potential performance gains.

An important advantage for inter-system cooperation can be derived from the use of measurements and triggers. Use of positioning technologies for deriving precise location information can be well exploited for support of individual user needs related to QoS. Further, operators can improve the network management by being able to reduce the number of unnecessary handovers. The location of the HIS functionality depends on the architecture and system deployment. In relationship to the proposed cooperation architecture, the HIS can be implemented through a central approach where the HIS can be located in the CoopRRM or a distributed approach where the HIS entities are distributed between the CoopRRM and the SRRM.

Different QoS requirements for different applications require a flexible and scalable RAN. Flexibility can be introduced by use of location information Applications that can use the location of the UT could be: point-to-point navigation, emergency call handling, location-based handover, or location-based service provisioning. In order to cater for different QoS demands, the location information should also be scalable in

order to provide for the appropriate accuracy for the involved UTs with different profiles. Generally, the location estimation can be done within the UT using measurements and information sent by the location service support function, or within the location service support using measurements sent by the UT and/or BSs and relay nodes involved in the location estimation process. The best stand-alone based performance can be obtained by including timing measurements of the UT. In-band timing measurements in cellular network based positioning are usually based on TDOA measurements. These are based on the idea to find the starting point (TOA) of the incident OFDM signals to estimate distances between the UT and the BSs using the included pilot sequences. If the GNSS based positioning information is included (GPS, Galileo), the performance can be further improved.

The following Chapter 3 and Chapter 4 propose cooperative RRM algorithms in support of QoS, network management and user mobility. Those algorithms are based on the measurement strategies proposed and analysed in this Chapter.

References:

[1] E. Mino, A., Mihovska, et al., "D4.2 Impact of Cooperation Schemes between RANs," Deliverable 4.2, IST Project WINNER, February 2005.

[2] M., Lott, A., Mihovska, et al., "Cooperation of 4G Radio Networks with Legacy Systems," *Proc. of IST Mobile Summit 2005,* Dresden, Germany, June 2005.

[3] E. Mino, A., Mihovska, et al., " D 4.8.1 WINNER II Intramode and Intermode Cooperation Schemes Definition," Deliverable D4.8.1, IST Project WINNER II, June 2006.

[4] M. Lott, V. Sdralia, M. Pischella, D. Lugara, A. Mihovska, S. Ponnekanti, E. Tragos, E. Mino, "Cooperation Mechanisms for Efficient Resource Management between 4G and legacy RANs," *Wireless World Research Forum (WWRF), 13th meeting,* Seoul, Korea, March 2005.

[5] Release 99, www.3gpp.org/Releases/3GPP_R99-contents.doc

[6] www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_26/Docs/PDF/SP-040900.pdf

[7] F., Meago, "Common Radio Resource Management (CRRM)", COST273, May 2002.

[8] IST project WINNER II, Deliverable D6.13.8, Final System Concept, November 2007.

[9] P., Gelpi, A., Mihovska , A., Lazanakis , G., Karetsos, B., Hunt, J., Henriksson , P., Oillikainen, and L., Moretti, "Scenarios from the WINNER Project: Process and Initial Results," *Wireless World Research Forum (WWRF), 11th meeting,* Oslo, Norway, June 2004.

[10] P., Karamolegkos, E., Tragos, A., Mihovska, et al., "A Methodology for User Requirements Definition in the Wireless World," *Proc. of IST Mobile Summit 2006,* Mykonos, Greece, June 2006.

[11] E., Mino, A., Mihovska, et al., "D 4.1: Identification and Definition of Cooperation Schemes between RANs," Deliverable 4.1, IST Project WINNER, June 2004.

[12] 3GPP TS 25.215, 3GPP; Technical Specification Group Radio Access Network; Physical layer - Measurements (FDD), Release 6, V6.4.0, 2005-09.

[13] E., Mino, A., Mihovska, et al., D4.3, "Identification, Definition and Assessment of Cooperation Schemes between RANs," Deliverable 4.3 IST project WINNER, June 2005.

[14] E., Mino, A., Mihovska, et al., D4.4, "Impact of Cooperation Schemes between RANs—A Final Study," Deliverable 4.4 IST Project WINNER, November 2005.

[15] A., Mihovska, et al., "Policy-Based Mobility Management for Next generation Systems," *Proc. of IST Mobile Summit 2007,* Budapest, Hungary, July 2007.

[16] A., Mihovska, et al., "Requirements and Algorithms for Cooperation of Heterogeneous Radio Access Networks," accepted for publication in the Springer International Journal on Wireless Personal Communications (ID WIRE 391) 2008.

[17] A., Bondavalli, "Model-Based Validation Activities," IST Project CAUTION, October 2003.

[18] A., Mihovska, et al., "QoS Management in Heterogeneous Environments," *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

[19] A. Mihovska, et al., "Algorithms for QoS Management in Heterogeneous Environments," *Proc. of WPMC'06,* San Diego, California, September 2006.

[20] X., Fang and D., Ghosal, "Analyzing Packet Delay Across A GSM/GPRS Network", IEEE 2003

[21] S., Kyriazakos and G., Karetsos, Practical Radio Resource Management in Wireless Systems, Norwood MA: Artech House 2004.

[22] A. Mihovska, "Cognitive Ubiquitous Mobile Communications," *2nd CTIF Workshop*, Aalborg, Denmark, May 2007.

[23] Mitola, J. III, Cognitive Radio Architecture, Wiley Publishers, 2006.

[24] A., Mihovska, et al., "Cooperative Radio Resource Management for Heterogeneous Networks," Chapter in the book on *Cooperative Wireless Communications,* to be published by Auerbach Publications, CRC Press, Taylor&Francis Group in July 2008.

[25] H., T., Nguyen, and E., A., Walker, *A First Course in Fuzzy Logic,* Chapman & Hall/CRC, Taylor & Francis Group, 2006.

[26] A., Konar, Computational Intelligence: Principles, Techniques and Applications, Springer-Verlag 2005: Berlin-Heidelberg.

[27] D., Dubois and H., Prade, "Fuzzy Sets and Systems: Theory and Applications," Academic Press, October 1980.

[28] A., Mihovska; H., Laitinen, and P., Eggers, "Location and Time Aware Multi-System Mobile Network," Proc. of Mobile Location Workshop'03, Aalborg, Denmark, May 2003.

[29] http://www.terabeam.com/support/calculations/antenna-downtilt.php.

[30] RECOMMENDATION ITU-R M.1645, "Framework and Overall Objectives of the Future Development of IMT 2000 and Systems Beyond IMT 2000," At www.itu.int.

[31] netlab18.cis.nctu.edu.tw/html/wlan_course/powerpoint/802.11f%20-%20IAPP.pdf.

[32] A.-G. Acx, A. Mihovska, et al., "D1.3 Final Usage Scenarios," Deliverable 1.3, IST 2003-507581 Project WINNER, at www.ist-winner.org.

[33] D. Tse, and P. Viswanath, *Fundamentals of Wireless Communications,* Cambridge University Press 2005.

[34] R. Prasad, W. Mohr, and W. Konhäuser, Third Generation Mobile Communication Systems, Artech House 2000.

[35] M., Cheung and J., W., Mark, "Resource Allocation in Wireless Networks Based on Joint Packet/Call Levels QoS Constraints," in Proc. of IEEE Global Telecommunications Conference (GLOBECOM '00), San Francisco, California, November–December 2000, Vol. 1, pp. 271–275.

[36] I., Ramachandran and S., Roy, "On the Impact of Clear Channel Assessment on the MAC Performance," *Proc. Of GLOBECOM,* San Francisco, California, November 2006.

[37] H.-H., Liu, J.-L.C., Wu, and W.-Y., Chen, "New Frame-Based Network Allocation Vector for 802.11b Multirate WLANs," *Proc. Of IEE Communications,* Volume 149, No. 3, June 2002.

[38] W., Ye, and J., Heidemann, " Medium Access Control in Wireless Sensor Networks, "Report, October 2003 at www.isi.edu/~weiye/pub/isi-tr-580.pdf.

[39] J., Kowalski, US Patent 20060046688, "Medium Sensing Histogram for WLAN Resource Reporting," February 2006.

[40] S. Black, "IEEE P802.11 Wireless LANs" Comment Resolution, March 2004.

[41] G., Landi, "Properties of the Centre of Gravity Algorithm," *Proc of Como Communications,* October 2003.

[42] W.-I., Kim et al., "Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks," IEEE 2007, pp. 1509-1512

[43] F., Gustafsson and F., Gunnarsson, "Mobile Positioning Using Wireless Networks," *IEEE Signal Processing Magazine,* July 2005, Vol. 22, No. 4.

[44] IST project WINNER, Deliverable 6.13.7, "WINNER Test Scenarios and Calibration Case Issues," December 2006 at www.ist-winner.org.

[45] B., W., Parkinson and J., J., Spilker Jr., "Global Positioning System: Theory and Applications, Volume 1," *Progress in Astronautics and Aeronautics,* Volume 163, 1996.

[46] E. Mino, A. Mihovska, et al., IST-4-027756 WINNER II D4.8.3 "Integration of cooperation on WINNER II System Concept," November 2007.

Chapter 3

Cooperative RRM for Handover

This Chapter proposes and evaluates RRM algorithms for inter-and intra-system handover. The proposed algorithms are assessed based on the proposed in Chapter 2 measurements strategies and follow the adopted in Chapter 2 framework for cooperative RRM.

Further to the realization of RRM techniques for handover, the location of the RRM functions is also studied. A combined centralised and distributed approach is proposed. Therefore, the proposed RRM algorithms are made consistent with the specifics of the RAN architectures of the investigated IMT-A reference system (see Figure 2-1) and the RANs of the legacy systems. The location of the RRM functions within the network architecture is an essential issue and can affect the performance if causing significant signalling and delays. In a centralised architecture, a central entity monitors and makes decisions regarding the allocation of resources and the user terminal (UT) has a minimal participation. In a distributed RRM architecture, the decision entities for each RRM function are located to different nodes, including the UT. A hybrid approach is also proposed, and there the decision levels of the same RRM functionality that can be active at different timescales are allocated to different nodes. The impact of the proposed algorithms on the proposed cooperation architecture is also studied in this Chapter.

This Chapter is organised as follows. Section 3.1 proposes and evaluates RRM algorithms for inter-system handover. The assumed scenario is for inter-system handover between an IMT-A candidate system and an UMTS system. The assessment investigates the effect of the network load and load thresholds on the process of inter-system handover for different numbers of UTs. Section 3.2 proposes RRM algorithms for intra-system handover. The proposed intra-system handover algorithms include inter-mode and intra-mode algorithms as part of the generic and specific RRM

algorithms of the proposed RRM framework, correspondingly (see Chapter 1 and Chapter 2).

3.1 Inter-System Handover

The scenario for inter-system handover algorithm assumes interworking between two RANs. RAN1 is an IMT-A candidate system operating in short-range mode (i.e., served by BS_{LA}) and RAN2 is an UMTS system. In general, the proposed scenario is in accordance with the one shown in Figure 1-2 and the protocol reference scenario shown in Figure 1-4. The inter-system handover procedure is considered at Layer 2 (MAC level). The goal is to satisfy the UT requirements by choosing the appropriate serving system. For this purpose, the inter-system handover algorithm is defined based on *coverage* and *load* criteria. The UT performance indicators [1] will have impact on the variation of the *load* threshold that triggers the handover from RAN2 to RAN1 (called *UMTS_HO_Load_thr*). The performance indicator is either the throughput or the sojourn time on each RAN. If the objective is to guarantee the accessibility to real-time services over UMTS, then the sojourn time on UMTS can be chosen as a quality indicator. Otherwise, the overall throughput is considered.

The following inter-system handover policy is studied: UTs are transferred to RAN1 as soon as *coverage* and *load* conditions are satisfied. The goal is to maximize their throughput. In order to apply the handover policy, a threshold is defined on the UMTS load that triggers handovers from RAN2 to RAN1. This threshold is denoted as *UMTS_HO_Load_thr*. If the *UMTS_HO_Load_thr* is **high**, UTs are still attached to UMTS until the system overloads.

If *UMTS_HO_Load_thr* is **low**, handovers to RAN1 become almost immediate. The advantage is that the UTs profit from higher throughputs on the RAN1.

The algorithm is generic and could be applied in both directions: from UMTS to RAN1 and from RAN1 to UMTS. Hereafter, the algorithm is described for handover from RAN2 to RAN1 in order to keep things clear and accurate. The following handover algorithm aggregates the proposed *coverage* and *load* criteria:

For each test period, for each UT:
**Find out the best received cell of the active set: $cell_i$.*
**Find out the cell with the lowest load: $cell_j$ among the cells of the active sets.*
** Find the BSla the best received by the UT among the neighbouring BSla.*
IF** ((Power received from $cell_i$ < UMTS_coverage_threshold) **OR** (Load of $cell_j$>UMTS_ load_threshold) **THEN
**Trigger measurements on RAN1*

At measurements end:
IF** ((Power received from $cell_i$ < UMTS_coverage_threshold) **OR** (Load of $cell_j$>UMTS_ load_threshold) **THEN

**Create the list of target BSla verifying the threshold conditions on coverage and load.*
**Find the target BSla the best received by the UT among the aforementioned list.*
**Trigger a handover to this BSla.*

The used coverage metric is *CPICH Ec/N0* (the received energy per chip divided by the power density in the band) [1] for UMTS and the signal strength for RAN1. They are both reported by the UT. The load on UMTS is defined as the ratio of the downlink power to the maximal downlink power:

$$Cell\ Load = 100 * \frac{Downlink\ Power}{Downlink\ Maximum\ Power} \quad \textbf{(3-1);}$$

The load on RAN1 is defined as the BS_{LA} buffer occupation or:

$$BS_{LA}\ Load = 100 * \frac{Queue\ Size}{Maximum\ Queue\ Size} \quad \textbf{(3-2)}$$

The overall downlink throughput is the throughput perceived by the UT over the whole activity time, on both UMTS and RAN1. It is computed as the data volume received from both RANs divided by the total downlink activity time:

$$Overall\ DL\ bit\ rate = \frac{D_{UMTS} + D_{LA}}{t_{total}} \quad \textbf{(3-3)}$$

For the evaluation of the performance of the proposed algorithm, three central and 12 neighbouring UMTS cells are considered. Within each of the three central cells, there are three LA cells. All UTs move at 3 kmph. UTs connect to UMTS first. They request a downlink radio access bearer (RAB) of 384 Kbps and an uplink RAB of 128 Kbps on UMTS. The maximum data rate of a BS_{LA} is 2 Mbps. The traffic model for each UT consists of a downlink FTP service (100% get). The FTP server sends files of 400 Kb with an inter-request time of 100s.

Simulations are performed for different values of *UMTS_HO_Load_thr* within the following set of values: {90, 80, 70, 60, 50, 40, 30, 20, 15}. The number of UTs varies from 5 to 25 at a step of 5. *UMTS_HO_Load_thr* has an impact on the variation on the overall throughput and the time spent on UMTS or different numbers of UTs. The handover policy consists in maximizing the overall throughput by triggering handovers from UMTS to RAN1. For this purpose, the *UMTS_HO_Load_thr* is decreased. Figure 3-1 shows the overall downlink throughput as a function of *UMTS_HO_Load_thr* for different numbers of UTs.

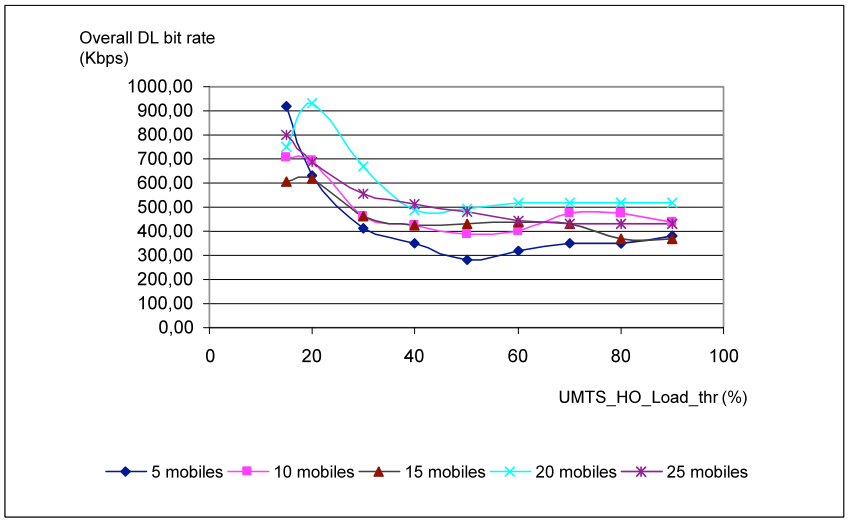


Figure 3-1 Overall load-based downlink throughput for different numbers of UTs.

There are two phases in the overall throughput evolution graph. During the first phase, the throughput is almost constant when *UMTS_HO_Load_thr* has low values. During, the second phase, the low value of *UMTS_HO_Load_thr* implies immediately the increase of the number of load-dependent handovers from UMTS to RAN1 and consequently the overall throughput increases. Figure 3-2 shows the ratio of load-dependent handovers from UMTS to RAN1 (e.g., IMT-A LA) versus *UMTS_HO_Load_thr* for different numbers of UTs. For the same threshold value, the throughput increases when the number of UTs goes up from 5 to 20 UTs. This is due to the fact that the more UTs are on UMTS, the faster the handover threshold is reached.

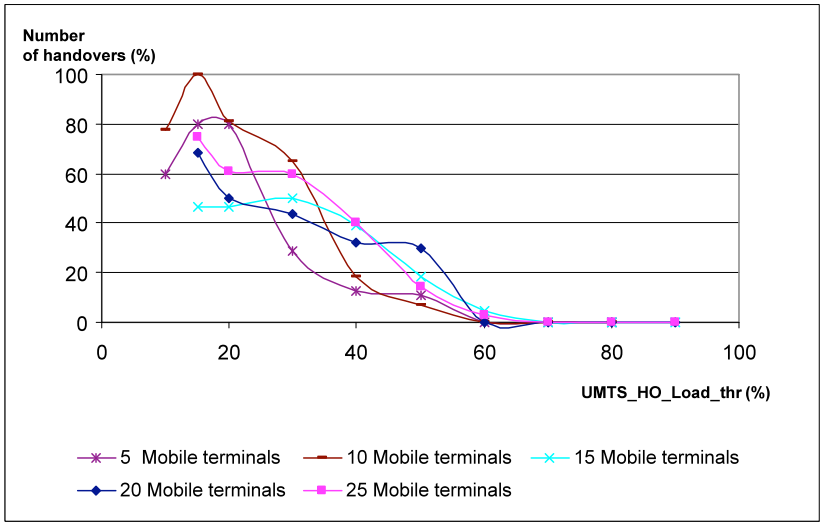


Figure 3-2 Percentage of load-based handovers from UMTS to IMT-A LA for different numbers of UTs.

Consequently, the handovers to RAN1 are frequent and the overall throughput increases. When the number of UTs reaches 20 to 25, the throughput slows down. The reason is that the BS_{LA} capacity is shared by more users, which leads to an overall throughput decrease. Then, for 25 UTs, the throughput increases again. In this case, some UTs are blocked due to the admission control and load control algorithms.

Figure 3-3 shows the ratio of sojourn time spent on UMTS versus *UMTS_HO_Load_thr* for different numbers of UTs. For a ratio of total duration on UMTS for 5 to 25 UTs, the sojourn time on UMTS increases with *UMTS_HO_Load_thr*, which is justified. For values of the *UMTS_HO_Load_thr* between 15 and 40 or 50 (according to the number of UTs), the sojourn time on UMTS increases rapidly and then becomes almost constant.

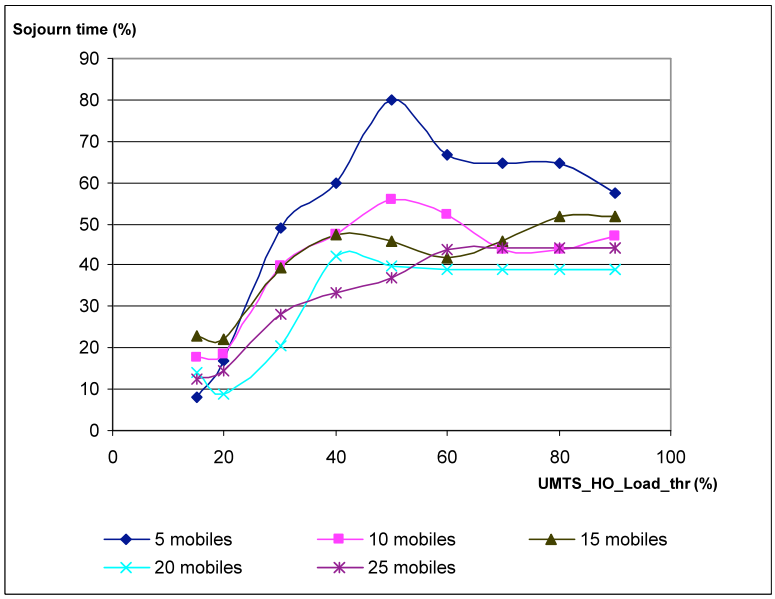


Figure 3-3 Load-based sojourn time spent on the network for different numbers of UTs .

Use of thresholds together with cooperative RRM algorithms is complex especially in a multi-system context. For each system, many algorithm parameters should be fixed: selection, access control and mobility algorithms thresholds. These different thresholds are correlated and should be set jointly. Moreover, the inter-system mobility algorithm parameters must be correlated with the RRM thresholds of each system.

3.2 Intra- System Handover

This section proposes handover algorithms for the cooperation between RRM entities of the same system as proposed for IMT-Advanced candidate systems [2]. In principle, the

concepts described for the cooperation between different RANs were extended to enable the cooperation inside the same RAN. The higher layer triggers are expected to be activated either by BS calculations on the cell status, or by information sent by the monitoring entities [3]. In this context the proposed general handover algorithm in Figure 2-2 is applicable also for support of intra-system handover. Intra-system handover will involve generic and specific RRM algorithms. Because the BS of IMT-Advanced candidate systems will serve different deployments [2], [4], two specific intra-system handover cases can be considered, namely: (1) when the handover takes place between BS serving the same deployment (*intra-mode handover*), and (2) when the handover takes place between BS serving different deployment areas (*inter-mode handover*). Complete details about the 'mode' concept are available in [2].

A handover process can be triggered by periodic measurements and by a higher layer trigger (e.g., cell load), then the UT requests to the network elements information on the possible cells of the same mode or different modes, or different RANs. Depending on the type of handover; intra-mode, inter-mode or inter-system, this information will be provided by a specific entity: the BSs, GW/SRRM or CoopRRM, correspondingly. In particular, inter-mode decision will be advised by the following entities:

- BS_{WA}/BS_{MA} will be deciding the handover between a LA and WA/MA;
- GW/SRRM will be deciding the handover between WA and MA.

If we follow a self-organized and partially distributed approach the intra-mode handover decision could be taken by the BSs /UTs of the same mode, in a similar way to the current 802.11 standards, i.e., without a central entity. The BSs, of the same mode in the same deployment zone, could use a protocol to exchange control messages between them, in a similar way that the 802.11 access points (APs) use the *Inter-Access Point Protocol (IAPP),* to give a continuous coverage in support of terminal mobility [5].

3.2.1 Intra-Mode Handover

There are three possibilities for intra-mode handover: between BSs of the same deployment type, between RNs and between RN and BSs of the same deployment. This type of handover includes the intra-cell handover where the user remains in the same mode (e.g., the change of frequency in the same cell) and the inter-cell handover between cells of the same mode. The basic trigger for inter-cell handover is the received

signal strength, but also, the load of the neighbouring cells, congestion situations, increased interference, the location of the user, etc. The intra-mode handover (between RNs and/or BSs), for example, could be triggered when the received signal strength (RSS) is below a fixed specified minimum value.

In the active state, when a data flow is requested by the UT or the network, this data flow is mapped to a service class. In [6], [7], [8] 18 service classes were proposed as exemplary for an IMT-A candidate systems. These service classes could be served by WA, MA or LA. In the case when the UT would be served by an adequate deployment scenario, as default, the UT will try to handover to a neighbouring cell of the same mode when an intra-mode trigger is activated. Only in the case when there are no cells of the same mode available, or when some specific inter-mode triggers are activated, the UT will handover to other mode or even another RAN (e.g., increase or decrease of UT velocity). The group of triggers were described in Chapter 2.

Figure 3-4 proposes the actions for the intra-mode algorithm, assuming that two different cells (current and target cells) can provide the QoS requirements related to the service requested by the user. The handover is based on the criteria of coverage and load as proposed in Chapter 2.

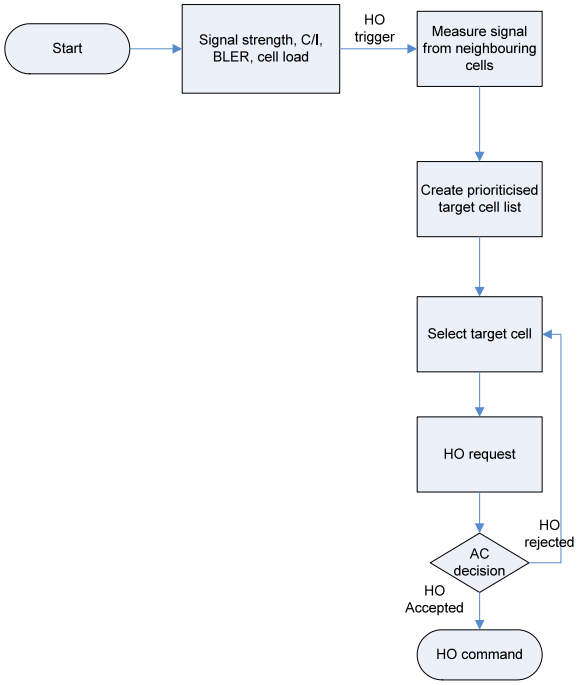


Figure 3-4 Intra-mode handover algorithm.

It should be noted that the threshold for the load criteria can be absolute or relative, if it is possible to compare directly the load. The criteria to choose the users (and their number) that perform the handover can be *services-based* criteria (e.g.,

speech users perform handover first, then another service, etc.), or *resources-based* criteria (i.e., the users consuming a lot of resources are transferred first, etc.) or *users-based criteria* (i.e., low-priority users are forced to handover first). In Figure 3-4 the UT performs periodic measurements and, when a trigger for intra-mode handover is activated, the signals from the neighboring cells lists are measured and the target cells are ranked. The admission control algorithm decides whether to accept or reject the user to the new cell; if the admission decision is positive the user performs the handover. Otherwise, it selects another cell from the target list and so on, until either the user is accepted in a cell or the user is rejected. The associated signalling with the intra-mode handover is shown in Figure 3-5.

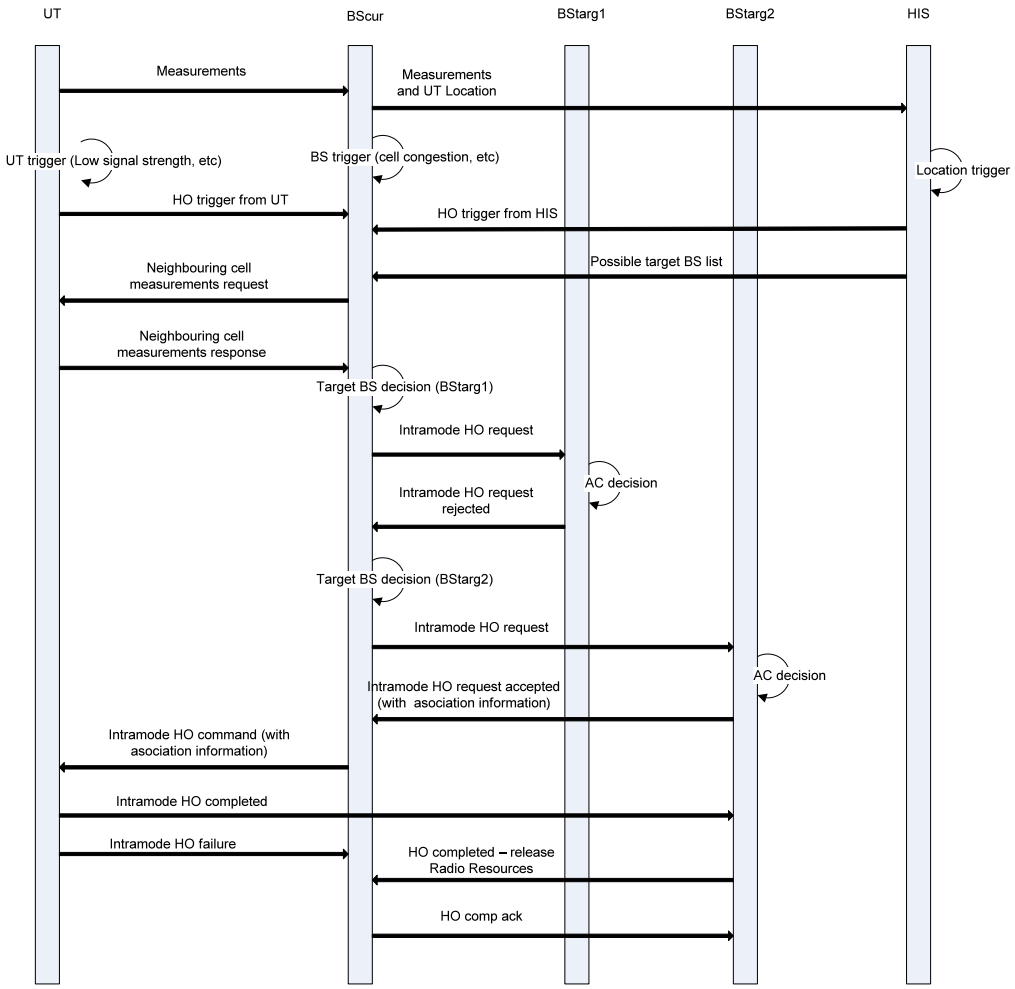


Figure 3-5 Signaling during intra-mode handover.

Upon receiving an intra-mode handover trigger, the current BS (BS_{cur}) gets the list of the neighbouring cells from the HIS and sends it to the UT. Now the UT knows, for which cells to measure the signal strength it receives and sends the measurements back to the BS_{cur}. The BS_{cur} then makes a list of the possible target cells and sends the

handover request to the new target BS (BS_{targ1}). Then the AC on that BS is activated. If the admission is rejected then the BS_{cur} sends the handover (HO) request to the next target BS (BS_{targ2}). When the AC accepts the handover, the BS_{targ2} sends the HO request acceptance message to the BS_{cur}, which sends the HO command to the UT. Then the UT sends the HO completed message to the target BS to request radio resources and the BS_{targ} response and sends also a HO completed message to the BS_{cur} to release the radio resources of the UT. The BS_{cur} acknowledges and releases the radio resources and the handover is completed.

3.2.2 Inter-Mode Handover

An inter-mode handover includes the handover from a BS_{WA} of a WA deployment scenario (or BS_{MA} of a metropolitan area deployment) to a BS_{MA} (or a BS_{LA} of a LA deployment scenario) and vice versa, correspondingly [9]. The intra-mode handover is a sub-set of the inter-mode handover. This is shown in Figure 3-6.

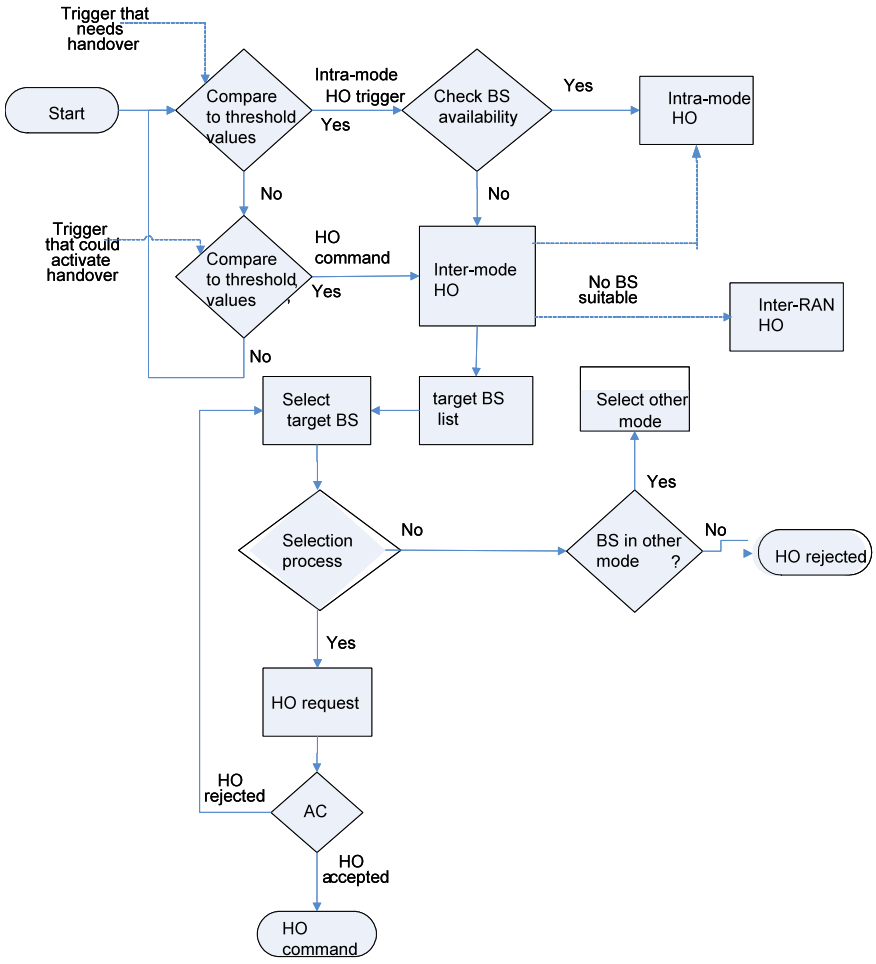


Figure 3-6 Inter-mode handover flow chart including the relationship to the intra-mode handover.

Inter-mode handover can be activated by the need for higher data rate services (e.g., handover from the WA to LA) or UT velocity (e.g., handover from the LA to the WA). After the initial selection of the service class and the associated modes to an user data flow, the UT will be maintained in this mode, until some changes would occur in the UT environment or data flow. When no cell is available then the cell of other RANs would be checked, in this case the cell selecting process in the legacy RAN will be similar to the inter-mode handover process.

Inter-mode handover could be initiated by the UT or the BS to which the user is initially connected. In this process, the BS_{WA} is the BS that coordinates the handover procedure. The BS_{WA} receives the request from the current BS for an inter-mode handover, gets the measurements from the UT and the current BS (i.e., $BS_{cur(LA)}$) and decides the list of modes and BSs of each mode that are suitable for the user. Then all the messages are exchanged via the BS_{WA} in order to complete the handover.

Inter-mode handover is triggered by low signal strength and quality (BER) and cell congestion triggers. There are specific triggers that directly activate inter-mode handover as, for example, new services request/release and velocity changes.

When there is an inter-mode handover trigger, the algorithm tries to find the best suitable mode for the user to handover to. This decision is based on several criteria which were analyzed above and also on the trigger that requested the handover. The target modes (if there is more than one suitable) are listed and ordered by preference according to the above criteria. For the selected mode a list of target cells is created and it is checked with the AC to which cell the user can be admitted. The associated signalling to inter-mode handover is shown in Figure 3-7. In this case the handover is from a BS_{WA} (or a BS_{MA}) to a BS_{MA} (or a BS_{LA}) and vice versa, correspondingly. The BS_{WA} is coordinating the handover procedure. The BS_{WA} receives the request from the current BS for an inter-mode handover, gets the measurements from the UT and the BS_{cur} and decides the list of modes and BS of each mode that are suitable for the user. Then all the messages are exchanged via the BS_{WA} in order to complete the handover. Taking into account that the BS_{MA} can control several RNs, it is foreseen that the BS_{MA} could control also several BS_{LA}. Therefore the BS_{MA} will have the same control functionality as the BS_{WA}. This is in essence a decentralised hierarchical approach to intra-system interworking. Centralised and decentralised approaches are playing an important role in the research proposed in later chapters.

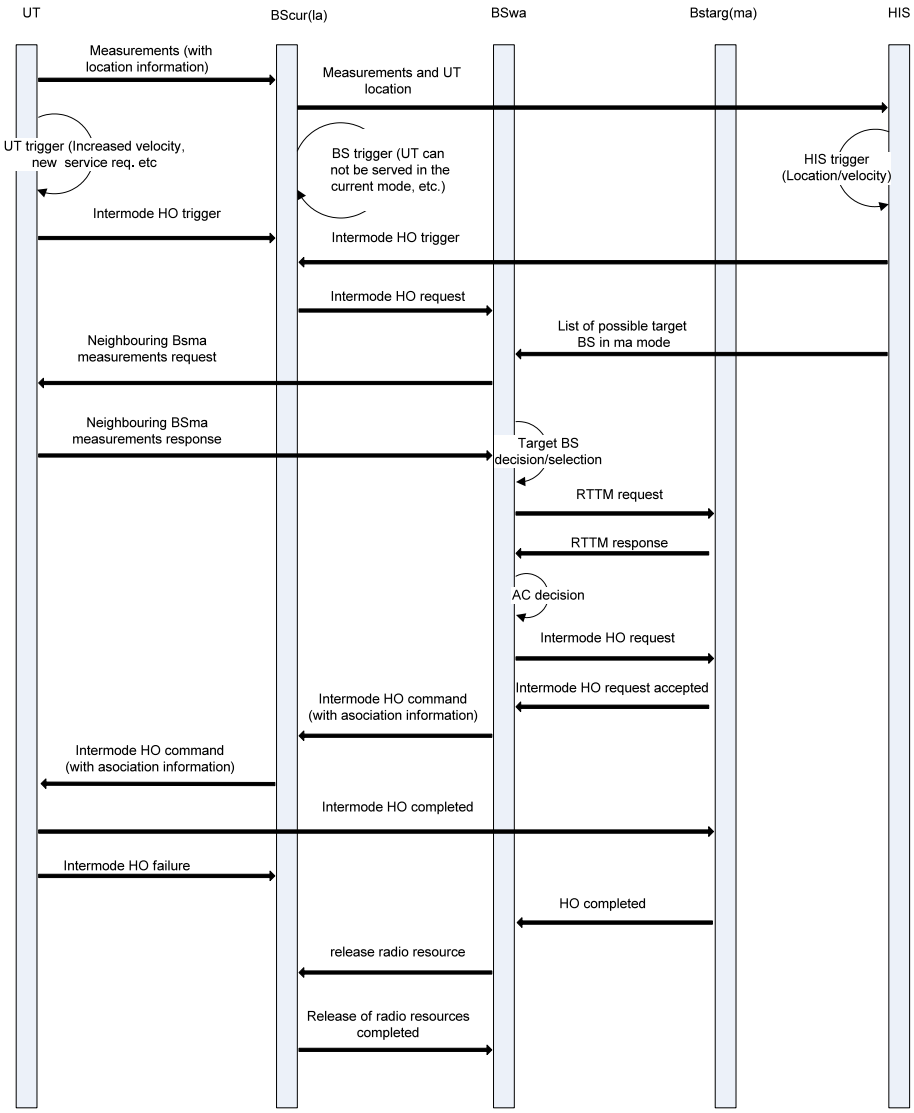


Figure 3-7 Signaling during inter-mode handover.

3.2.3 Hierarchical Control Architecture for Intra-System Handover

A decentralized and hierarchical mode control architecture is proposed for the intra-system RRM algorithms, which would improve the efficiency of the handovers. In such an approach, the inter-mode handover decision is taken by the $BS_{WA/MA}$ that controls several BS_{LA}. In this approach the GW node will be limited to coordinate a set of BS_{WA} or BS_{MA}. With other words, the GW would govern over a pool of BS, which would also decrease the need for handovers when a UT is roaming within such a pool of BS. This is shown in Figure 3-8.

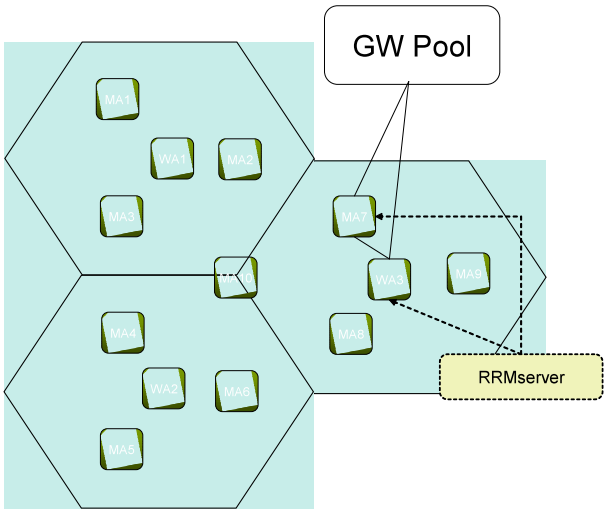


Figure 3-8 Concept of a hierarchical control for generic RRM.

The benefit of the proposed hierarchical control architecture in Figure 3-8 is that mode generic control plane functions that concern the coordination of the different modes/BSs could be moved to the BS_{WA} and BS_{MA}, making them responsible for the control and allocation of resources per WA cell including all BS_{LA} that fall within its coverage.

A requirement for such an approach would be the definition of a communication link between the BS_{WA} and BS_{LA}, this link could be either wired or wireless (e.g. part of the WA mode interface).

3.2.3.1 Hierarchical Control Architecture Involving Several GWs

Several GWs can be employed for optimizing the GW pool capacity. The GW is an anchor point for external routing, and also is the bridge between the UT and the operator services and Internet, through the I_G interface. Adding or removing GWs can help balance the load between GWs. Finally, it provides redundancy, that is, in the case of a GW failure, the users can be handed over to any other GWs in the pool, and at the same time load balancing between GWs can easily be achieved. Load balancing strategies are proposed in Chapter 6.

The GW association will be preserved even when a user handovers to a BS that is controlled by a different GW, belonging to the same pool. In that case, a change of the IP address is not necessary. This is the basic idea for the proposed pool of GWs that are connected to the BSs through a routing function that enables the process described above. This concept is shown in Figure 3-9.

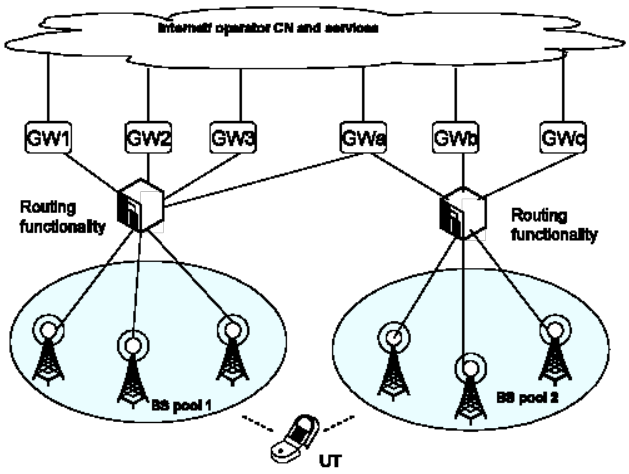


Figure 3-9 Pool of GWs communicating through a routing function.

3.2.3.1 Communication between BS during intra-system interworking

The intra-system interworking is based on two connection solutions, i.e., through the backbone network provided by the BSs connections and, in addition, over the air interface between overlapping BSs. The BS-BS interface is viewed as beneficial for the communication between BSs that belong to two pools of GWs in the context of opportunistic communications [10]. The latter functionality can be included as an optional and on demand functionality[1]. This is a proposal in line with the concept of the Hierarchical Cell Structures (HCS) [11]. The only difference is the proposed over-the-air communication between BSs. The proposed scenario is shown in Figure 3-10.

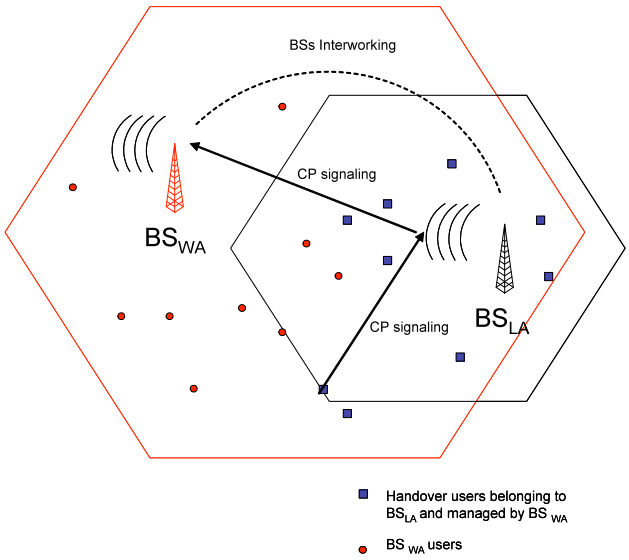


Figure 3-10 Proposed scenario for interworking between BSs.

[1] The most efficient way to for inter-BS communication in support of RRM actions is planned as a follow up work of this thesis and is explained in Chapter 7. Chapter 4 presents a preliminary proposal for efficient BS-BS communication in support of user context transfer during handover.

The process is described for the message exchange scenario shown in Figure 3-7.

The UT requests the approximated inter-mode measurements from the BS_{LA} after the activation of a pre-trigger. For example, if a PHY-layer trigger is activated based on intra-mode measurements performed by the UT then the rule *bit error rate (BER)>pre_trigger_limit* initiates the exchange of measurements with the BS_{LA} while *BER>trigger_limit>pre_trigger_limit* necessitates the handover to WA).

After the UT has sent its intra-mode measurements to the BS_{LA}, it would request the approximate measurements for the BS_{WA}. These measurements, however, will not be used to request a handover unless a trigger that necessitates the handover is activated (see Chapter 2). The UT might need to perform inter-mode measurements itself in the case when a trigger that necessitates the handover occurs before a pre-trigger (e.g., fast increase in the UT velocity) or the last pre-trigger occurred some time ago (the duration of this timeout will be based on the air interface specifications) and therefore the inter-mode measurements are not considered valid. Based on the BS_{WA} cell and BS_{LA} cell information, the BS_{LA} decides whether to accept, to decline or to queue the UT handover request. To inform a UT that is within the range of a MA/WA cell, a beacon signal should be sent periodically by the BS_{WA}. Upon the receipt of the beacon, the UT could perform any further measurements preparing for an inter-mode handover. Therefore, such a handover is initiated by the BS rather than the UT.

In Figure 3-11, handovers are performed based on the proposed intra-system RRM algorithm and are triggered by a value of the throughput, which is measured in Mbps. The scenario is the same of Figure 3-10 when two BS_{LA} are in the vicinity of the BS_{WA}.

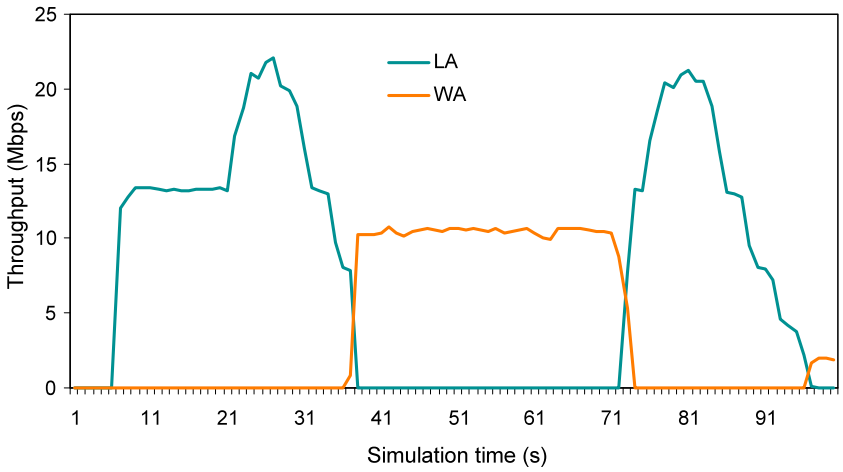


Figure 3-11 Intra-system handover algorithm triggered by residual throughput.

When the throughput achievable from a LA connection decreases, the UT will handover to WA and similarly, when it comes close to the BS_{LA} it would handover to it.

A beacon will be transmitted on the broadcast channel common for all system modes and, therefore, the UTs will be able to receive cell information in whichever mode they operate at any time. This could be implemented by having the BS_{LA} transmitting on the broadcast frequency of the BS_{WA} (but adding complexity to the BS_{LA} because a second transmitter must be added) or by switching between the modes. Furthermore, whenever the BS_{LA} will transmit, only UTs within the LA cell will receive the beacon. Because all BS (BS_{WA} and BS_{LA}) within a WA-cell coverage will transmit information on the broadcast channel, synchronisation between the BSs is required. It is proposed here that each BS is allocated periodically the same timeslots of the frame [9].

In particular, the BS_{WA} will be responsible for allocating these timeslots flexibly. For example, assuming a MAC frame of k timeslots and i BS_{LA}, the frame can be equally divided to (i+1) parts, which will be allocated by the BS_{WA} (k> (i+1)). In another implementation, the BS_{WA} will allocate a whole frame per BS_{LA} in a Round Robin fashion [12]. Finally, the BS_{WA} could dynamically allocate frames or timeslots according to requirements for broadcasting information. The broadcast beacon should at least include information on the current LA cell identification number. Further to the common broadcast channel, a second broadcast channel just for the MA/WA cell where the beacon is also transmitted for the terminals that are already at MA/WA mode is needed. As the path from the BS to the UT might not be direct, the information message should also include any nodes identification number which can be added on the beacon as each node propagates it. Other useful information that could be included would be the functionality of each RN, conventional or cooperative. This will assist the UT to identify the number of hops as well as what relay functionality is available. Therefore, the UT might receive as many beacons as the number of links from the BS. The advantage of this method is that the UT will know how many paths (this is in effect routing information) are available to it as well as which is the best link (between last relay and UT), (e.g., by calculating the BER of the beacon message). This information and the number/type of hops between the BS_{LA} and the UT could be a parameter for deciding on which path to send/receive information. In addition the information form neighbouring cell lists can assist the handover process (see Section 2.2.1 of Chapter 2).

3.3 Impact on Cooperation Architecture

To enable the implementation of the proposed RRM framework, it is proposed that the CoopRRM has interfaces with other CoopRRM of the same or different operators.

Further, it is proposed that the logical functionality of the CoopRRM is divided in a *common* part (RRM-g) and a *specific* part (RRM-s) for each RAN with the common part containing the functionalities common to all RANs. The RRM-g provides a common interface towards upper layer functions/protocols. The specific part handles the specific details of each RAN.

In summary, it is proposed that the SRRM module (i.e., RRM Server and GW) includes the following functionalities:

- Receive real time traffic measurement;
- Calculate KPIs;
- Forward alarms to CoopRRM;
- Provide status to CoopRRM on demand;
- Enable RRM-s.

The alarm and status information provide the driven force for the handling of the cooperation between the RANs. This information is passed from the SRRM as structured format-based information and would be referred to as radio resource control (RRC) signalling. Explicit RRM functions use the RRC signalling and implement a set of suitable functions to support intelligent admission of calls and sessions. They control the distribution and the association of traffic, power and the variances of those, for an optimized usage of radio resources and maximized system capacity. Control refers to the decision made by the measuring station or remote entity to adjust the radio resources based on the reported measurements, or to activate the RRM functions. Further, the RRC communicates the adjustments to the logical entities using standardized primitives.

The RRC includes measurements, exchange and control of radio resource-related indicators and commands between the RAN and UTs. The measurements are the determining values of standardized radio resource indicators that measure or assist in estimation of the available (and potentially available) radio resources.

In the control plane, the I_{GB} establishes a flow context in a BS by a GW. The BSs informs the GWs about the user mobility (e.g., handover and paging updates). Similarly, the BS provides means to forward the paging messages to a certain UT.

In order to provide for scalability and flexibility of the proposed RRM architecture, it is proposed that a combined centralised and distributed approach is used at different layers of the framework. The dotted lines in Figure 2-1 showed the signalling of the optional RRM entities, which when activated would act in a centralised manner. Inter-system cooperation would always be performed in a centralised way,

whereas, intra-system cooperation will be handled in a distributed manner for low traffic loads. Therefore, the I_{GB} interface is defined as the logical interface between the GW and BS (mode-generic and mode-specific RRM for low loads). The I_{RRM} interface will be active for medium to high loads. Use of the optional entities is CAPEX efficient for real network deployment. The analysis of the degree of efficiency, however, has not been included here. In terms of delivering QoS, such an approach can guarantee that for dense areas with users demanding high QoS, a centralised approach will be able to guarantee the QoS.

The RRM server can give potential gains provided by the centralised JRRM through the interfaces between the RRM server and the BSs. The system capacity gain obtained from the deployment of the RRM server is in principle the enlargement of the number of operational servers from the queuing model viewpoint, which therefore results in a higher trunking gain. The potential gains from alternatively allocating the resources to call units among the interworking coexisting BSs and the realisable load balancing effect are studied further in Chapter 4 and Chapter 5.

From this point on, the definition/understanding of centralized and decentralized RRM is the following.

• *Decentralized RRM* means that decisions are taken at cell level by the BS, independently from other cells. Information from other cells, (e.g., measurements), can be considered in the decision process.

• *Centralized RRM* means decisions are at least influenced (even though not necessarily made) by RRM entities located higher in the network hierarchy and consider information reported from other cells/RANs.

3.3.1 Inter-Function Cooperation for a Hybrid Approach

It is proposed to optimise the interworking among the proposed RRM functions by a hybrid approach in support of intra-system interworking. This optimization is shown in Figure 3-12. Because congestion control itself is not an explicit function which evokes explicit air interface updates, it is classified as a network function. Explicit RRM functions use the RRC signaling and implement a set of suitable functions to support intelligent admission of calls and sessions. In a situation of high load to medium load, the congestion function may trigger (network-triggered) handovers by shifting some selected users to another cell/mode or frequency in order to avoid the congested state. To handle the handover process, the RRM functions need the explicit RRC messages, which are termed as '*explicit function through the air'*.

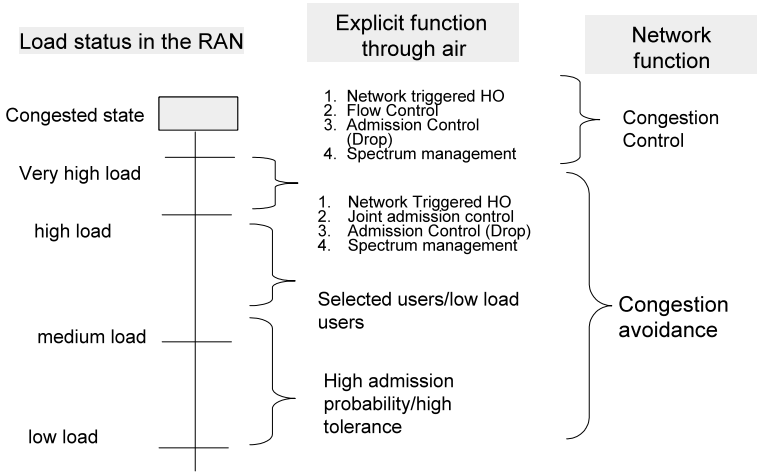


Figure 3-12 Optimised interworking among the RRM functions.

The numbers attributed to the functions follow the sequence of functions execution for a congested state or a scenario of very high load.

To demonstrate the interworking among the RRM function in the three different approaches and map them to the required RRC signalling, the centralised, distributed and hybrid approaches are proposed and analysed for intra-system handover. This is further detailed in Chapter 5.

3.4 Conclusions

This chapter proposed cooperative RRM mechanisms for inter-system and intra-system interworking. It was shown that cooperation is required on three levels, supported by cooperative, mode generic and mode specific RRM mechanisms. Such an approach has benefits for emerging communication systems, because the proposed framework ensures the coexistence with legacy systems. The proposed RRM framework is based on the proposal for a flat distributed RAN envisioned for next generation systems and already approved for LTE. Further it ensures that active inter-connections between relevant RRM entities are maintained at the desired by the cooperation level. The inter-node connectivity provides interworking, which in turn provides potential performance gains.

An important advantage for inter-system cooperation can be derived from the use of measurements and triggers. Use of positioning technologies for deriving precise location information can be well exploited for support of individual user needs related to QoS. Further, operators can improve the network management by being able to reduce the number of unnecessary handovers. The location of the HIS functionality

depends on the architecture and system deployment. In relationship to the proposed cooperation architecture, the HIS can be implemented through a central approach where the HIS can be located in the CoopRRM or a distributed approach where the HIS entities are distributed between the CoopRRM and the SRRM.

Inter-system cooperation is best coordinated by a centralised approach and by an entity located externally to the RANs. Inter-system RRM mechanisms including congestion, admission and load control can reduce the percentage of blocked and dropped users even for scenarios of high traffic load. The advantage of the proposed here inter-system cooperation mechanisms is mainly that they allow for decision making based on generic factors, such as the load of the system, while considering the individual user and service characteristics. In a heterogeneous mobile environment, it is difficult to rely on exact mathematical descriptions or on prior knowledge for all of the processes and interactions in each system and therefore, this approach allows for scalability of the framework. Further, a large portion of the information related to the system performance and radio resources allocation is to be found in the users and applications of that system. With other words, the proposed mechanisms allow for data from the edge of the system to be combined with data from other parts of the system thus understanding the complete sequence of events.

Cooperation can be assisted further by introducing a combined centralised and distributed (i.e. hybrid) approach to cooperation for support of intra-system cooperation and for introducing scalability of the architecture, which is important for network migration. A policy-based management framework based on this combined approach is proposed in Chapter 5, whereas a novel multi-stage admission control algorithm, also based ont eh combined RRM approach is proposed in Chapter 6.

References:

[1] 3GPP TS 25.215, 3GPP; Technical Specification Group Radio Access Network; Physical layer - Measurements (FDD), Release 6, V6.4.0, 2005-09.

[2] IST project WINNER II, Deliverable D6.13.8, "Final System Concept," November 2007, at http://www.ist-winner.org.

[3] E. Mino, A., Mihovska, et al., " D 4.8.1 WINNER II Intramode and Intermode Cooperation Schemes Definition," Deliverable D4.8.1, IST Project WINNER II, June 2006.

[4] RECOMMENDATION ITU-R M.1645, "Framework and Overall Objectives of the Future Development of IMT 2000 and Systems Beyond IMT 2000," At www.itu.int.

[5] netlab18.cis.nctu.edu.tw/html/wlan_course/powerpoint/802.11f%20-%20IAPP.pdf.

[6] P., Gelpi, A., Mihovska , A., Lazanakis , G., Karetsos, B., Hunt, J., Henriksson , P., Oillikainen, and L., Moretti, "Scenarios from the WINNER Project: Process and Initial Results," *Wireless World Research Forum (WWRF), 11th meeting,* Oslo, Norway, June 2004.

[7] P., Karamolegkos, E., Tragos, A., Mihovska, et al., "A Methodology for User Requirements Definition in the Wireless World," *Proc. of IST Mobile Summit 2006,* Mykonos, Greece, June 2006.

[8] A.-G. Acx, A. Mihovska, et al., "D1.3 Final Usage Scenarios," Deliverable 1.3, IST 2003-507581 Project WINNER, at www.ist-winner.org.

[9] A., Mihovska, et al., "Requirements and Algorithms for Cooperation of Heterogeneous Radio Access Networks," in the *Springer International Journal on Wireless Personal Communications* (ID WIRE 391), DOI: 10.1007/s11277-008-9586-y, August 2008.

[10] D. Tse, and P. Viswanath, *Fundamentals of Wireless Communications,* Cambridge University Press 2005.

[11] R. Prasad, W. Mohr, and W. Konhäuser, Third Generation Mobile Communication Systems, Artech House 2000.

[12] M., Cheung and J., W., Mark, "Resource Allocation in Wireless Networks Based on Joint Packet/Call Levels QoS Constraints," in Proc. of IEEE Global Telecommunications Conference (GLOBECOM '00), San Francisco, California, November–December 2000, Vol. 1, pp. 271–275.

[13] E. Mino, A., Mihovska, et al., "D4.2 Impact of Cooperation Schemes between RANs," Deliverable 4.2, IST Project WINNER, February 2005.

[14] M., Lott, A., Mihovska, et al., "Cooperation of 4G Radio Networks with Legacy Systems," *Proc. of IST Mobile Summit 2005,* Dresden, Germany, June 2005.

[15] M. Lott, V. Sdralia, M. Pischella, D. Lugara, A. Mihovska, S. Ponnekanti, E. Tragos, E. Mino, "Cooperation Mechanisms for Efficient Resource Management between 4G and legacy RANs," *Wireless World Research Forum (WWRF), 13th meeting,* Seoul, Korea, March 2005.

[16] Release 99, www.3gpp.org/Releases/3GPP_R99-contents.doc

[17] www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_26/Docs/PDF/SP-040900.pdf

[18] F., Meago, "Common Radio Resource Management (CRRM)", COST273, May 2002.

[19] E., Mino, A., Mihovska, et al., "D 4.1: Identification and Definition of Cooperation Schemes between RANs," Deliverable 4.1, IST Project WINNER, June 2004.

[20] E., Mino, A., Mihovska, et al., D4.3, "Identification, Definition and Assessment of Cooperation Schemes between RANs," Deliverable 4.3 IST project WINNER, June 2005.

[21] E., Mino, A., Mihovska, et al., D4.4, "Impact of Cooperation Schemes between RANs—A Final Study," Deliverable 4.4 IST Project WINNER, November 2005.

[22] A., Mihovska, et al., "Policy-Based Mobility Management for Next generation Systems," *Proc. of IST Mobile Summit 2007,* Budapest, Hungary, July 2007.

[23] A., Bondavalli, "Model-Based Validation Activities," IST Project CAUTION, October 2003.

[24] A., Mihovska, et al., "QoS Management in Heterogeneous Environments," *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

[25] A. Mihovska, et al., "Algorithms for QoS Management in Heterogeneous Environments," *Proc. of WPMC'06,* San Diego, California, September 2006.

[26] X., Fang and D., Ghosal, "Analyzing Packet Delay Across A GSM/GPRS Network", IEEE 2003

[27] S., Kyriazakos and G., Karetsos, Practical Radio Resource Management in Wireless Systems, Norwood MA: Artech House 2004.

[28] A. Mihovska, "Cognitive Ubiquitous Mobile Communications," *2nd CTIF Workshop*, Aalborg, Denmark, May 2007.

[29] Mitola, J. III, Cognitive Radio Architecture, Wiley Publishers, 2006.

[30] A., Mihovska, et al., "Cooperative Radio Resource Management for Heterogeneous Networks," Chapter in the book on *Cooperative Wireless Communications,* to be published by Auerbach Publications, CRC Press, Taylor&Francis Group in July 2008.

[31] H., T., Nguyen, and E., A., Walker, *A First Course in Fuzzy Logic,* Chapman & Hall/CRC, Taylor & Francis Group, 2006.

[32] A., Konar, Computational Intelligence: Principles, Techniques and Applications, Springer-Verlag 2005: Berlin-Heidelberg.

[33] D., Dubois and H., Prade, "Fuzzy Sets and Systems: Theory and Applications," Academic Press, October 1980.

[34] A., Mihovska; H., Laitinen, and P., Eggers, "Location and Time Aware Multi-System Mobile Network," Proc. of Mobile Location Workshop'03, Aalborg, Denmark, May 2003.

[35] http://www.terabeam.com/support/calculations/antenna-downtilt.php.

[36] I., Ramachandran and S., Roy, "On the Impact of Clear Channel Assessment on the MAC Performance," *Proc. Of GLOBECOM,* San Francisco, California, November 2006.

[37] H.-H., Liu, J.-L.C., Wu, and W.-Y., Chen, "New Frame-Based Network Allocation Vector for 802.11b Multirate WLANs," *Proc. Of IEE Communications,* Volume 149, No. 3, June 2002.

[38] W., Ye, and J., Heidemann, " Medium Access Control in Wireless Sensor Networks, "Report, October 2003 at www.isi.edu/~weiye/pub/isi-tr-580.pdf.

[39] J., Kowalski, US Patent 20060046688, "Medium Sensing Histogram for WLAN Resource Reporting," February 2006.

[40] S. Black, "IEEE P802.11 Wireless LANs" Comment Resolution, March 2004.

[41] G., Landi, "Properties of the Centre of Gravity Algorithm," *Proc of Como Communications,* October 2003.

[42] W.-I., Kim et al., "Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks," IEEE 2007, pp. 1509-1512

[43] F., Gustafsson and F., Gunnarsson, "Mobile Positioning Using Wireless Networks," *IEEE Signal Processing Magazine,* July 2005, Vol. 22, No. 4.

[44] B., W., Parkinson and J., J., Spilker Jr., "Global Positioning System: Theory and Applications, Volume 1," *Progress in Astronautics and Aeronautics,* Volume 163, 1996.

Chapter 4

Cooperative RRM Algorithms for Congestion, Admission and Load Control

This Chapter proposes the protocols in support of cooperative RRM algorithms for admission, congestion and load control. The algorithms then are assessed for a scenario of inter-system interworking for medium to high load (i.e., busy hour). In the cooperative RRM framework, these algorithms are closely interacting with each other to provide for QoS management including network capacity optimisation. Chapter 2 (see Figure 2-7) showed the variations of the delay depending on the chosen congestion threshold and exercised load. It was shown that excessive loading leads to a higher delay value and consequently into drop of the throughput. Admission control (AC) in this context refers to a functionality, which grants or rejects user requests based on the network resource availability, which in turn depends on the pre-defined load/congestion thresholds and the instantaneous load values. In broad terms, AC limits the access to some resource such that the load on that resource remains limited. Different load situations will affect the network functionalities differently. Careful handling of the load and admission control strategies can also help reduce the number of unnecessary handovers. Therefore, the proposed congestion, admission and load control algorithms are also distinguished as *cooperative, mode generic* and *mode specific*. The scope of this research work extends only to the cooperative and mode generic algorithms. Mode specific algorithms are primarily targeting the transmission control functionality at lower layers, and are therefore, not in the scope of this research work.

This Chapter is organised as follows. Section 4.1 proposes a cooperative admission control algorithm and analyses its relationship to the system loads. This analysis is used later for the assessment model. Section 4.2 proposes a cooperative congestion control algorithm. Again, it is analysed in relationship to the system loads. Section 4.3 assesses the proposed algorithms in terms of achievable system capacity and QoS. Section 4.4 analyses the impact of the proposed algorithms on the RRM architecture. Section 4.5 concludes the Chapter.

4.1 Cooperative Admission Control

Figure 4-1 shows the proposed AC algorithm for cooperation between different RANs.

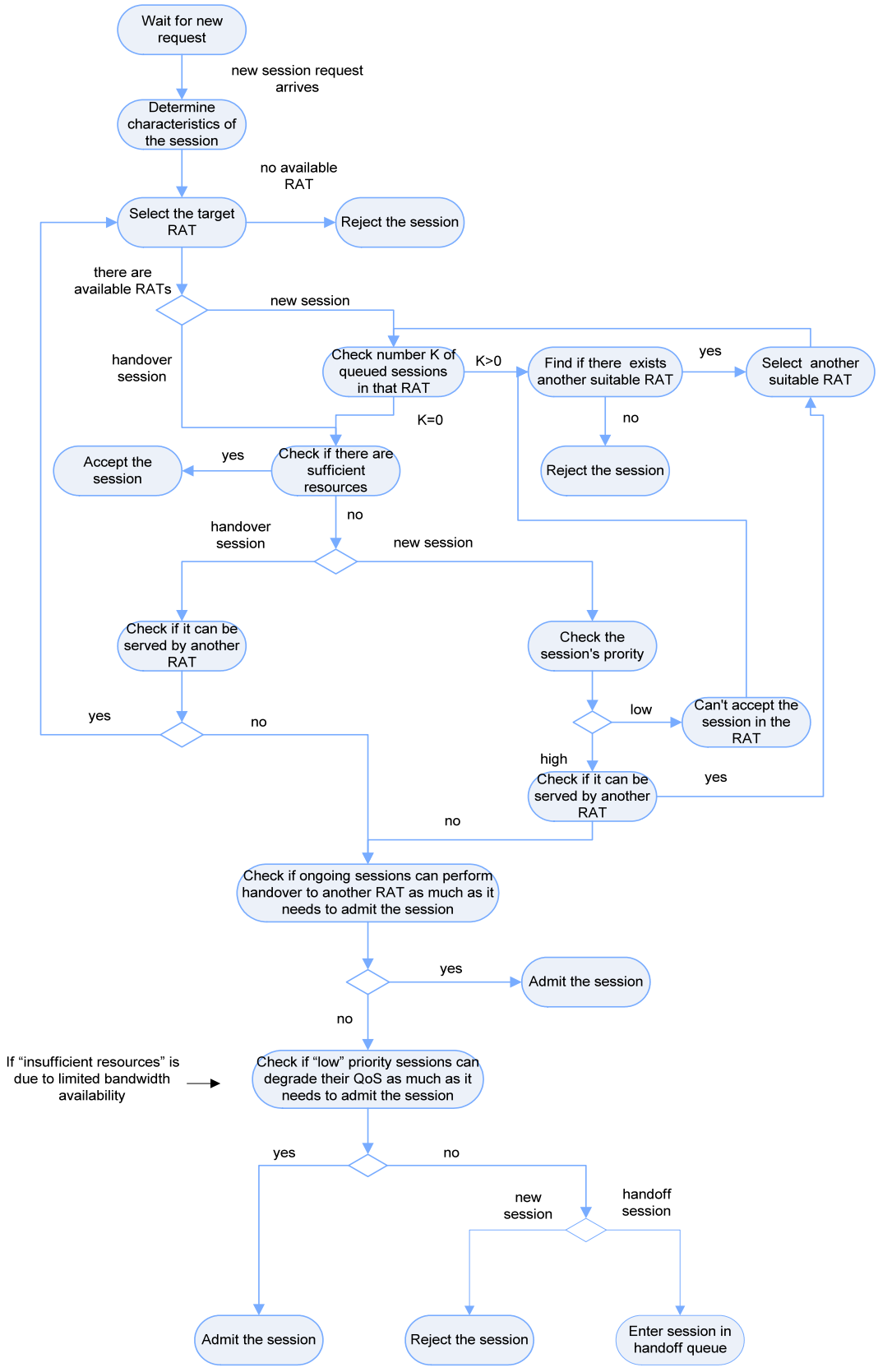


Figure 4-1 AC algorithm for cooperation between RANs.

It is based on a centralised approach with the main functionalities residing at the CoopRRM entity. The algorithm is triggered when a new session request arrives at the AC entity. A *"new session request"* can be a request from a new or a handover session.

When a new request arrives at the AC entity, then the algorithm will be executed in order to decide whether there is a RAN/RAT that can meet the session requirements and if the session can be served by that RAN/RAT.

First the characteristics of the session are determined. At this step the algorithm checks the requirements of the session, by means of resources.

The session declares its type, bandwidth requirement, delay sensitivity, and so forth. The session is then matched assigned a priority according to the service class it belongs to, the user profile and so forth. In order to select an appropriate RAN, the algorithm makes a list of the candidate serving RANs and candidate cells for each network.

The lists contain the candidate networks and cells capable of providing the requested session service and they are ordered in a way that fulfills the service requirements in each network. For example, in GSM/GPRS candidate cells are the ones that provide coverage at the point where the user is located, that means that the BCCH Rx_{lev} value is greater than a defined threshold [1]. In UMTS, at the point where the user is located, the E_b/N_o of the cell must be above a specific value and the transmission power requested to the user by the power control loop has to be lower than the maximum allowed value for that user [2]. In WLAN the coverage is reached when the received power is above a minimum value [3].

The decision to admit or reject a user is based on a calculation of the load and comparing it to a predefined threshold. This was also described in Chapter 2. Each of the RANs calculates the load in a way reflecting the specifics of the interwokring RAN architecture. This action is performed by the SRRM entity belonging to each RAN.

The algorithm in Figure 4-1 will maximize the number of admitted or in-session traffic sources supported over the RANs, while guaranteeing their QoS requirements and ensuring that the new connection does not affect the QoS of the ongoing connections. The decisions to accept or reject a new connection are based on the characteristics of the RAN. The algorithm assumes that degradation of the QoS of some users is acceptable provided that their QoS requirements are not violated.

For the assessment of the algorithm in Figure 4-1 the scenario assumes an IMT-A candidate RAN interworking with legacy RANs, such as GSM/GPRS, UMTS, and WLAN 802.11.

Users are admitted based on how this would affect the total load of the network; or the load must stay under the pre-defined load threshold (L_{th}) value:

$$Load < L_{th} \quad \textbf{(4-1);}$$

where L_{th} is different for the different types of networks.

As a basis, the load is defined from the average number of UTs requesting service or the average arrival rate, λ [users/sec], and the average time a UT requires service or the holding time, T [sec]. A channel kept busy for one hour is defined as one Erlang [4]. If the average arrival rate during a short time interval t is λt, then assuming Poisson distribution of service requests, the probability $P(n, t)$ for n calls to arrive at time interval t is given by Equation (4-2):

$$P(n,t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad \textbf{(4-2)}$$

Assuming μ to be the service rate, the probability of each call to terminate at time t is then μt. Thus, the probability that a given call requires service for a time t is given by:

$$S(t) = 1 - e^{-\mu t} \quad \textbf{(4-3)}$$

In GSM/GPRS, the load can be computed based on the number of occupied time slots (*TSused*) compared to the total number of time slots (*TSmax)* in a cell and considering the number of reserved slots. This means that a new user can be admitted if not all the time slots are occupied by other users.

The load of the GSM/GPRS network is computed according to Equation 4-4:

$$Load_{GSM} = 100 * \frac{TS_{used}}{TS_{\max}} \quad \textbf{(4-4);}$$

where

$$TS_{used} = TS_{RT} + TS_{NRT} \quad \textbf{(4-5);}$$

The final expression for the load of the GSM network is:

$$Load_{GSM} = 100 * \frac{TS_{RT} + TS_{NRT}}{TS_{MAX}} \quad \textbf{(4-6).}$$

If all the timeslots are occupied, then $TS_{used} = TS_{max}$. This means that in the situation where all the timeslots are occupied by users, the load of the network is 100%. If this situation occurs, no more users can be admitted to the network and a congestion control algorithm is activated to resolve the situation.

For the UMTS network the load is computed both for the uplink and downlink separately [2]. A user is admitted in the network if both uplink and downlink criteria are met. In the uplink, the criteria will be related to the received interference. During the planning phase of an UMTS network, the operator defines a maximum load for the network, given by η_{max}. If there are K admitted users in the system already, then another request for admission should meet the Equation 4-7:

$$\eta_{UL} + \Delta\eta \le \eta_{max} \quad \textbf{(4-7);}$$

where,

$$\eta_{UL} = \frac{P_R + \chi}{P_R + \chi + P_N} \quad , \quad \text{and} \quad \Delta\eta = \frac{1}{\frac{W}{v_{K+1} \cdot \left(\frac{E_b}{N_o}\right)_{K+1} R_{b,K+1}} + 1}$$

η_{UL} is the load of all the already admitted users in the uplink; P_R is the received power from the users in the cell and χ is the interference level coming from the neighboring cells. $\Delta\eta$ is the load increase that will be caused if the new user is admitted to the system. W is the chip rate, $(E_b / N_o)k+1$ is the target signal to noise ratio (SNR) of the new service, $R_{b,k+1}$ is the transmission rate of the new user and v_{k+1} is the activity factor of the new user's traffic source. If several service classes with different characteristics are assumed, then Equation (4-8) holds:

$$P_R = P_{R1} + P_{R2} + \ldots + P_{Rm} = \sum_{i=1}^{m} P_{Ri} \quad \textbf{(4-8);}$$

where P_{Ri} is the received power of the users of the service class i who are in the same cell as the new user.

Then a user is admitted in the network only when the load increase caused by the new connection does not make the current load of the network exceed the maximum load defined by the network operator. As a different criterion, one can consider the

interference level of the network [2]. The UL noise increase due to the new user admission is defined as:

$$\Delta N = \frac{I_{\text{int}ra} + I_{Inter} + P_N}{P_N} = \frac{I_{total}}{P_N} = \frac{1}{1-\eta} \quad \textbf{(4-9);}$$

where I_{intra} is the interference level caused by the users allocated in the same cell, I_{inter} is the interference level caused by users in the neighboring cells and P_{N} the noise power level. For assessing the AC algorithm of Figure 4-1, the I_{intra} level is defined to incorporate the number of service classes defined for the IMT-Advanced reference architecture [5]:

$$I_{\text{int}ra} = I_{R1} + I_{R2} + \ldots + I_{Rm} = \sum_{i=1}^{m} I_{Ri} \quad \textbf{(4-10);}$$

where I_{Ri} is the interference of the users of the service class i who are in the same cell as the new user. Thus, in order to check if the new user can be admitted or not in the network, the decision would be based on whether the following condition holds:

$$I_{total} + \Delta I \le I_{total,\,\max} \quad \textbf{(4-11);}$$

where I_{total} is obtained from the received total wideband power, P_{RX} , as in

$$I_{Total} = P_{RX} - P_u \quad \textbf{(4-12);}$$

P_{u} is the received power of the new user, and ΔI is the increase in the interference caused by the new user and defined by:

$$\Delta I = \frac{I_{total}}{1-\eta-\Delta\eta} \cdot \Delta\eta \quad \textbf{(4-13);}$$

with $\Delta\eta$ the load increase defined in Equation 4-7.

Thus, the uplink decision criteria are defined as in Equation 4-9 and Equation 4-11. The thresholds defined by Equation 4-7 and Equation 4-11 are not constant, and they depend on the type of service class. Users from different service classes will have different thresholds, so the system will be able to admit higher priority users above

lower priority users and depending on how they affect the thresholds. It is important that the thresholds are very well defined to avoid instability of the system [6].

In the DL the maximum transmitted power is shared among all the users located in the same cell. Also, the specific location of each user plays an important role and determines the amount of interference. The main criterion for the AC, therefore, is the transmitted power and not the load factor, as it was defined for the UL. The decision will be based on Equation 4-14:

$$P_{AV}(i) + \Delta P_T(i) \le P_T^*(i) \quad \textbf{(4-14);}$$

where,

$$P_{AV}(i) = \frac{\sum_{j=1}^{T} P_T(i-j)}{T} \quad \textbf{(4-15);}$$

P_{AV} is the average transmitted power during the last T frames, $\Delta P_{T(i)}$ is the power increase estimation caused by the new user and $P_{T(i)}$ is the threshold for the admission. An approximation of ΔP_T can be found from proposed algorithms [7], [8]. The calculation for ΔP_T adopted here is estimated with the power demand of previous users in a window of T frames:

$$\Delta P_T(i) = \frac{\sum_{j=1}^{T} \left(\frac{\sum_{k=1}^{n_{i-j}} P_{Tk}(i-j)}{n_{i-j}} \right)}{T} \quad \textbf{(4-16);}$$

n_{i-j} is the number of users transmitting in the $(i-j)$-th frame, T the averaging period (in frames) and P_{Ti} is the transmitted power to each user in the cell expressed by:

$$P_{Ti} \ge L_p(d_i) \frac{P_N + \chi_i + \rho \times \frac{P_T}{L_p(d_i)}}{\frac{SF_i}{\left(\frac{E_b}{N_o}\right)_i r} + \rho} \quad \textbf{(4-17);}$$

where P_T is the transmitted power, χ_i is the inter-cell interference to the user i, $L_p(di)$ is the path loss at distance d_i, r the coding rate and P_N the background noise. SF

compares the bit duration to the chip period and ρ is the orthogonal factor since orthogonal codes are used in the DL direction.

A "pessimistic" estimation of the ΔP_T value would assume the 90% of the required transmitted power per user or

$$\Delta P_T = P_{Ti}(90\% CDF) \quad \textbf{(4-18);}$$

where CDF is the cumulative distribution function derived from the probability density function (PDF) of the required transmitted powers to each user in the cell, P_T.

For the WLAN the decision is taken based on the user receiving power. The number of users that are being served by a WLAN access point (AP) and the amount and type of traffic of the radio interface play an important role in determining the load of the WLAN system [9].

In the 802.11e standard each service set has up to four different access categories with different priorities [10]. The method that is being proposed here takes into account the different priorities of the access categories and requires that each station measures the traffic condition (traffic load) on the wireless link. Two criteria are used here for making the decision about admitting a new user in the network. The first criterion is the *relative occupied bandwidth*, B_{occu}. In this method, the AC mechanism uses a time window in order to measure the amount of time used for transmission during a period T. This is the time when the wireless medium is busy, regardless of the fact whether the transmission has been successful or not. T is defined as follows:

$$T = \sum_{i=1}^{m} t_i \quad \textbf{(4-19);}$$

where t_i is the occupied time of the i-th transmission. Then the relative occupied bandwidth can be computed as follows:

$$B_{occu} = \frac{T_{busy}}{T} * 100 \quad \textbf{(4-20);}$$

The relative occupied bandwidth indicates the percentage of time that the wireless medium is busy (is being used). We define two thresholds here, B_{lo} and B_{up}.

If $B_{occu} > B_{up}$ then the wireless medium is in a congestion situation and the congestion control must take place. At this situation, no new users can be admitted in the network.

If $B_{occu} < B_{lo}$, there are free timeslots for new users. New users can be admitted to the network according to their priority.

If $B_{lo} \leq B_{occu} \leq B_{up}$ the wireless medium is close to congestion and only users with high priority are admitted to the network.

B_{occu} is computed at every period T and compared to the predefined thresholds. The algorithm could be easily implemented in the 802.11e network, because the enhanced distributed coordination function (EDCF) uses CSMA/CA medium access protocol, where a station has to sense the medium and check the network allocation vector (NAV) in order to see if the medium is idle for transmission of data [11]. The commonly used beacon interval of 802.11 could be used as the sampling period T.

The second criterion is the *average collision ratio,* R_c. In this method the AC mechanism uses a time window to measure the average collision ratio during a period T. The average collision ratio is defined as the number of collisions that have occurred during this period over the total number of transmissions (including retransmissions). The average collision ratio can be considered as a measure for the traffic load of the wireless medium. The average collision ratio, R_c, can be defined by:

$$R_c = \frac{N_c}{N_t} \qquad \textbf{(4-21);}$$

where N_c is the number of collisions in the period T and N_t is the total number of transmissions in that period.

Each station in the service set computes its own collision ratio R_c. Two thresholds, R_{lo} and R_{up} can be defined here (similar to the bandwidth). The following rules apply:

If $R_c > R_{up}$, the network is overloaded and the congestion control mechanism is activated. No new users are admitted in the network.

If $R_c < R_{lo}$, the network can receive new users without degradation of the performance.

If $R_{lo} \leq R_c \leq R_{up}$ the network is at an optimal state. At this situation no new users can be admitted to the network without degrading the QoS of the already admitted users. In this situation we allow only high priority users to be admitted to the network, by degrading the QoS of low priority already admitted users.

The 802.11e network already has the parameter of the total number of retransmissions. Even though it includes the retransmissions due to collision and the retransmissions because of received erroneous frames, the number of retransmissions could be a rough estimation of the collision rate (especially when the frame error rate is very small).Also here, the beacon interval of the 802.11 networks can be used as the period *T* [9].

The AC algorithm in Figure 4-1 is based on a very tight cooperation between the RANs in order to know at any time the characteristics of every legacy RAN. Therefore, the AC entity imposes requirements on the measurement entity. These requirements will be used as inputs for the AC algorithm in order to make the decisions about the session requests. The AC entity should be able to know at any specific time the condition of every RAN in terms of resources, such as occupied and available channels, number of connections per cell, load per cell, power and bandwidth availability of each cell etc. Inability to obtain this information would lead to wrong decisions and degradation of the QoS.

To assess the algorithm in Figure 4-1, a user generator was assumed that generates users with different characteristics and requesting sessions from the networks in specific locations. The users have the following characteristics:

userID, new/HO user, time_of_request, x_pos, y_pos, service_req, RAN_req, priority and call_duration

New users would want to establish a connection with one network and request admission for a new service, and handover users are the ones that would need to perform a handover. For the assessment it is considered that a new user would be accepted to the network. Based on the location of the user and taking into account the characteristics of the session the algorithm searches for a network and cells that are not overloaded and that can provide the user with the best QoS, either by selecting cells that are not overloaded or by performing several actions in order to gain the needed load in those cells (e.g., congestion control algorithm).

4.2 Cooperative Congestion Control

The cooperative congestion control based on the proposed AC is shown in Figure 4-2. A network is congested when the available resources are not sufficient to satisfy the experienced traffic load.

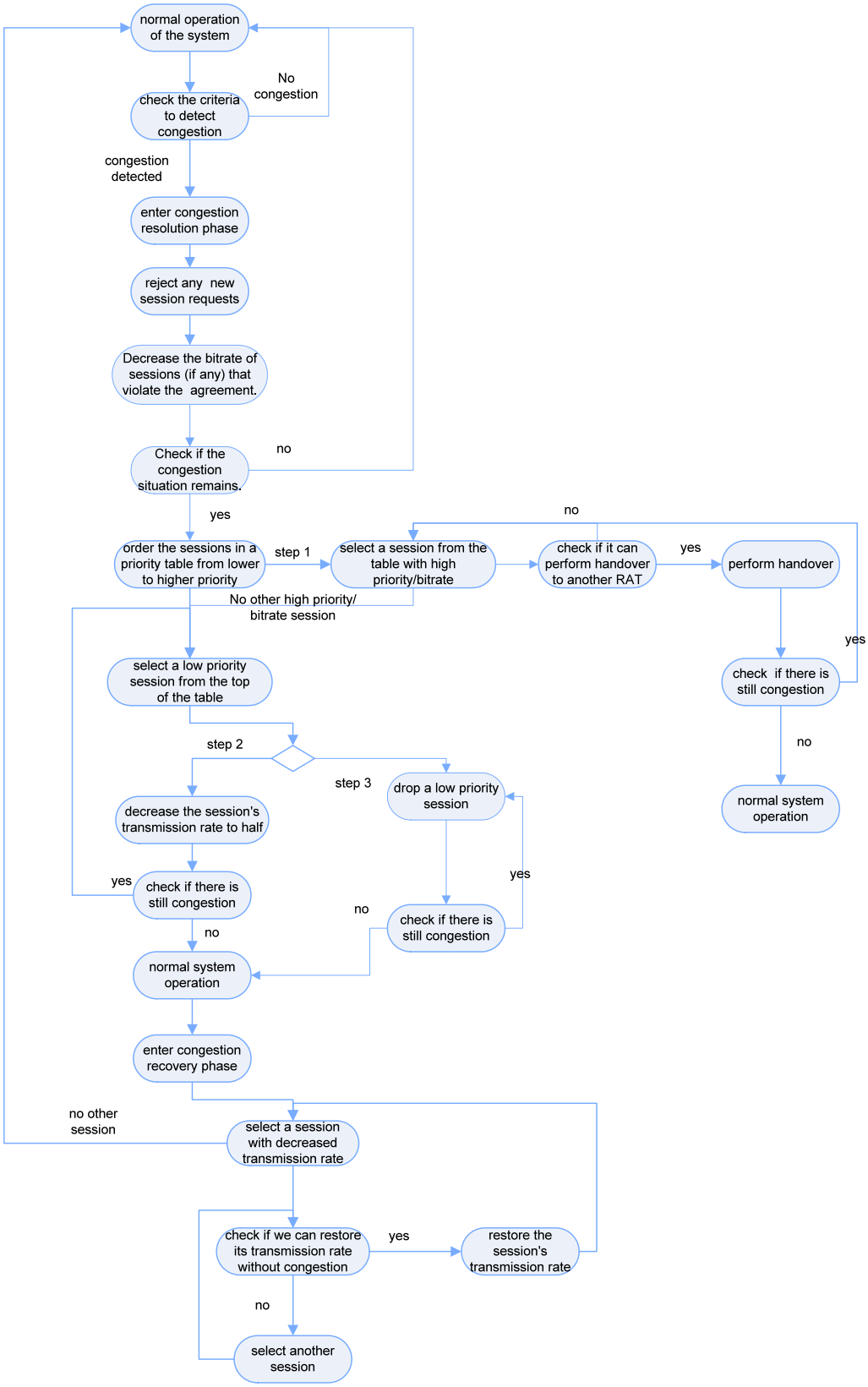


Figure 4-2 Congestion control algorithm for inter-system cooperation

There are two ways to induce congestion:

1. The network experiments a traffic overload that cannot be totally covered by the available resources, because the traffic rapidly increases inside a group of contiguous

cells. This is the case of an *emergency situation* (e.g., earthquake, major terrorist attack) or of an extraordinary accumulation of user requests because of special events (e.g., sport events, New Year's Eve);

2. An *outage* occurs. An outage is the unavailability of (part of) the network resources, typically because of malfunctions somewhere. The outage is distinguished as:

a) *Total outage* when the network is completely blocked and the on-going data transmissions are interrupted. This happens, for example, when the antenna of a cell is damaged and then no signals are sent or received;

b) *Partial outage* when only some of the resources are not available. This happens, for example, when some traffic channels are not properly working. The global service is still available but it operates in a degraded manner.

The above congestion situations in each RAN are solved by the specific for this RAN congestion control algorithms residing at the SRRM entities. For example, congestion in the UMTS network is solved based on the mechanism proposed in [12], and in the 802.11e system based on mechanisms proposed in [9], [11], [13]. For the reference IMT-A RAN adopted for the assessment, the congestion control mechanism monitors the network and if an overload situation occurs it would attempt to decrease the load of the network by performing several actions. One of these actions includes the activation of *a reactive load control algorithm* residing at the GW or SRRM entity, The reactive load control procedure is shown in Figure 4-3.

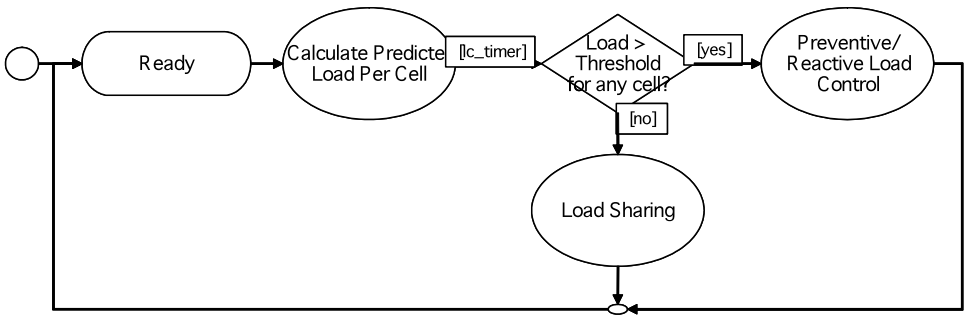


Figure 4-3 Reactive load control at GW and SRRM entity.

Figure 4-4 shows the proposed interworking between the load control algorithms residing at different levels of the RAN hierarchy.

Mode specific load control will reside at the BS entity based on thresholds indicative for the radio resource usage. A mode generic load control will also reside at the BS entity but will be active only at low network loads (below L_{th}) whereas for medium to high loads, the control algorithms residing at RRMServer (i.e., SRRM) and

GW will be activated. A novel multi-stage admission and load control algorithm based on the interworking introduced here is proposed in Chapter 6.

Different priorities can be assigned to the various steps of the load control to decide on, which action should be taken first in a given situation. These have not been assessed here.

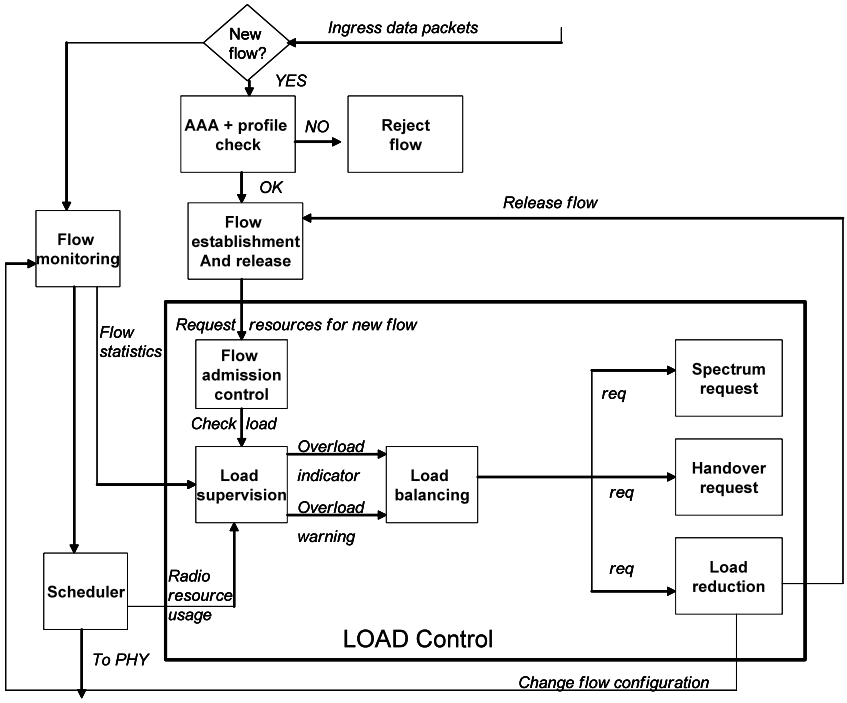


Figure 4-4 Reactive load control algorithm for the inter-system cooperation.

By admitting users automatically without AC, the load of the cells increases without controlling it. When the load reaches or exceeds a certain defined threshold then congestion control is triggered in order to decongest the cell/RAN. To reduce the complexity for the simulation scenario, the morphology of the cells was not considered, and square cells were assumed. A square-shaped cell is characterised by the topology shown in Figure 4-5 [4].

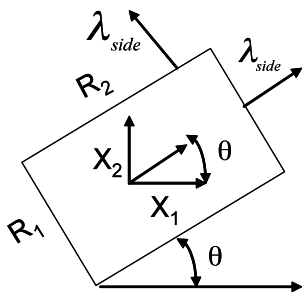


Figure 4-5 Topology of a square-shaped cell.

For example, the handover rate in a square-shaped cell can be calculated as follows. A handover can occur on one of the sides of the square (i.e., cell), R_1 or R_2. Then the handover rate λ_H can be defined as:

$$\lambda_H = R_1\left(X_1\cos\theta + X_2\sin\theta\right) + R_2\left(X_1\sin\theta + X_2\cos\theta\right) \quad \textbf{(4-22).}$$

If the area **A** = R_1 R_2 is assumed constant, then differentiation with respect to R_1 and R_2 gives:

$$R_1^{\,2} = A\frac{X_1\sin\theta + X_2\cos\theta}{X_1\cos\theta + X_2\sin\theta} \quad \textbf{and} \quad R_2^{\,2} = A\frac{X_1\cos\theta + X_2\sin\theta}{X_1\sin\theta + X_2\cos\theta} \quad \textbf{(4-23)}.$$

Then the total handover rate λ_H can be expressed as follows:

$$\lambda_H = 2\sqrt{A\left(X_1\cos\theta + X_2\sin\theta\right)\left(X_1\sin\theta + X_2\cos\theta\right)} \quad \textbf{(4-24);}$$

For θ = 0, λ_H is minimised.

4.3 Assessment Results

A total of 25 square-shaped cells were used for the simulations. The topology considers an IMT-A RAN and legacy RANs (i.e., GSM/GPRS, UMTS and WLAN 802.11b hotspots). The topology layout is shown in Figure 4-6.

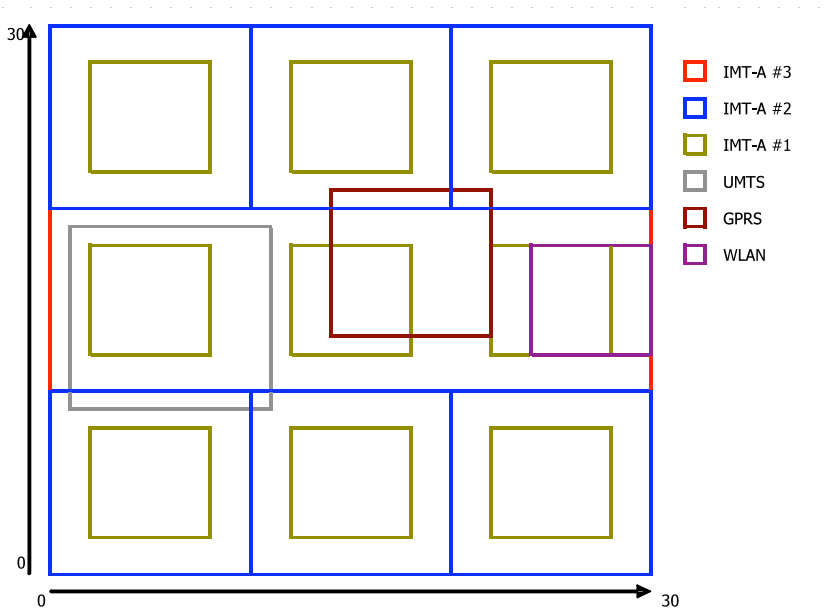


Figure 4-6 Topology for assessment of cooperative RRM.

The following cells have been used: 8 GPRS, 4 UMTS, 6 WLAN hotspots, 5 IMT-A LA and 1 IMT-A WA. The area of study is limited to the highest coverage area. This means that the whole area of study is entirely covered by the one IMT-A WA cell. The positioning of the cells is done by placing the upper left vertex of the square-shaped cell in the area of study. It is further assumed that all legacy RANs have the same capacity and that the IMT-A RAN has twice the capacity. To determine the total capacity C_{total} of the network, it has been assumed that a maximum of 5 000 users can be served within one hour. The total capacity value is calculated as in Equation 4-25:

$$C_{total} = \frac{N_{U\_\max} \sum_{i=1}^{N_{SC}} f_i DR_i TD_i}{FD} \qquad \textbf{(4-25);}$$

where $N_{\text{U\_max}}$ is the maximum number of users (i.e., 5 000); N_{SC} is the total number of service classes (i.e., 18); f_{i} is the penetration factor of the i^{th} service class; DR_{i} is the simulated data rate of the i^{th} service class, in Mbps; TD_{i} is the typical duration, or expected download time, of the i^{th} application, in seconds; FD is the full duration of the time interval, in seconds (i.e., 3600). The total capacity of the entire network was calculated as 2.625 Mbps. The load is calculated independently from Equation 2-4 for each cell. It is expressed as the occupation of the total capacity of the cell, and the load value is then comprised between 0 and 100. The calculation of the load for the IMT-A candidate RAN assumed here, takes into account only the bandwidth metric, in the data rate sense. For simplicity matters the same process has been applied to other RANs. This allows the algorithm to work with a generic formula. Moreover it provides load values that are coherent for the simulation, i.e., the load values for each RAN are comparable since they do not come from totally different calculations. The typical delay values were calculated by taking into account the method described in Chapter 2.

The simulations were based on the service classes defined in, [5], [15], [16], [17]. The service classes are shown again in Table 4-1 and Table 4-2. The requirements for each service class have been defined based on a thorough literature survey and public data from reports of the UMTS forum, 3GPP bodies, WiMAX forum, IST EU-funded projects and mapped to WINNER system requirements.

Table 4-1 Service Classes Characteristics

ID	description	Priority	Duration(sec) (min-max)		Data rate(kbps) (min – max)		BER (min – max)		Delay(msec) (min – max)	
SC1	Large files exchange	8	50MB	500MB	1000	50000	1,00E-06	1,00E-06	200	
SC2	High quality video streaming	6	300	600	2000	40000	1,00E-09	1,00E-09	200	
SC3	LAN access and file service	4	120	300	500	50000	1,00E-06	1,00E-06	100	200
SC4	Interactive ultra high media	1	120	500	1000	50000	1,00E-03	1,00E-06	20	100
SC5	Lightweight browsing	5	300	900	64	512	1,00E-06	1,00E-06	200	
SC6	Data and media telephony	2	60	120	64	512	1,00E-03	1,00E-06	100	200
SC7	Simple telephony and messaging	3	10	120	8	64	1,00E-03	1,00E-06	100	200
SC8	Multimedia messaging	7	5	15	8	64	1,00E-06	1,00E-09	200	

Table 4-2 Mapping of Service Classes to RANs

Class ID	Class name	GPRS	UMTS	WLAN	IMT-A
SC1	Large files exchange	-	-	-	X
SC2	high quality video streaming	-	-	-	X
SC3	LAN access and file service	-	-	x	X
SC4	interactive ultra high media	-	-	x	X
SC5	Lightweight browsing	-	x	x	X
SC6	data and media telephony	-	x	x	X
SC7	simple telephony and messaging	x	x	x	x
SC8	multimedia messaging	x	x	x	X

The assessment results are based on 8 service classes common to all RANs. The priorities are determined taking into account the needs in terms of interactivity (i.e., the maximum tolerable delay), and the resources in terms of throughput. Application priority will then refer to the priority of the service class that the application belongs to. Users running applications of high data rate would possibly require more resources and therefore, the service priority will be of a higher rank. Table 4-2 maps the service classes to the capabilities of the RANs assumed for the simulations. This allows for predicting, which networks will be able to serve a given user, and only those would be

checked by the algorithm. The most important parameter for the assessment results is the generated traffic.

The average duration of a call is computed from the duration defined for each session and from the size of the data that has to be exchanged according to the service class. These numbers are subjective and some of them extracted from the data rate that is defined for each class and the size of the data. From these and from the probability weight defined for each call, an average duration of the calls (i.e., the holding time at the BS) is computed as equal to 200 sec.

In the assessment results the average traffic generated in each cell is given by the equivalent throughput (see Figures 4-7 to 4-11). The values are a conversion of circuit switch Erlang into throughput based on the data summarised in Table 4-3. The Erlang as an expression of the channel capacity was introduced in relation to the average arrival rate earlier. Here the blocking probability [18] is defined based on Equation 4-1 to express the amount of offered traffic in Erlangs and the probability that an incoming call is being blocked. It is given by Equation 4-26:

$$P_B = \frac{A^N / N!}{\sum_{n=0}^{N} A^n / n!} \qquad \textbf{(4-26)},$$

where P_B is the blocking probability, A is the offered traffic in Erlangs, and N is the number of traffic channels available. One Erlang is one continuously used traffic channel, so during any given period (i.e., 3600s) if a user talks for half the time, there would be a generation 0.5E. If there are 10 users all talking for half the time, the total traffic load would be 5E.

The efficiency can be calculated as the ratio of the total amount of non-blocked traffic and the system capacity as defined by Equation 4-25.

The average equivalent throughput per cell for the scenario here is calculated by multiplying the Erlangs of each call by the throughput of the service class of the call with a distribution factor of 50%. The results are summarised in Table 4-3.

Table 4-3 Traffic Generated in a Cell

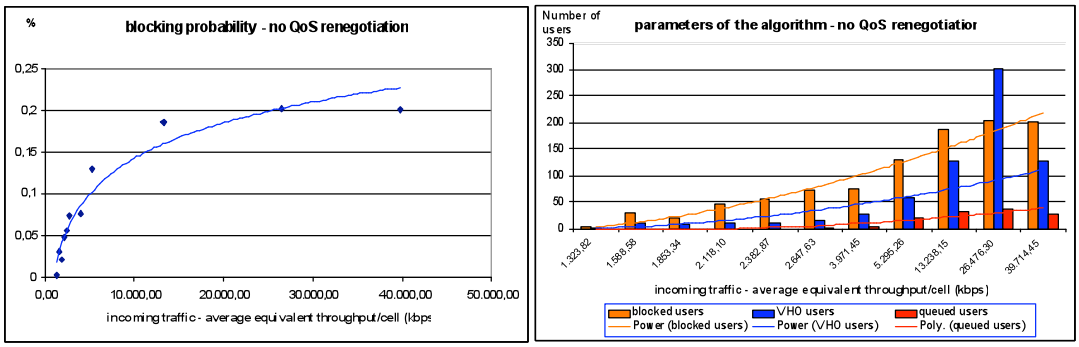
λ (incoming users/sec)	Average equivalent throughput/cell (kbps)	Average equivalent Erlangs/cell
0,5	1.323,82	4,14
0,6	1.588,58	4,97
0,7	1.853,34	5,80
0,8	2.118,10	6,63
0,9	2.382,87	7,46

1	2.647,63	8,29
1,5	3.971,45	12,43
2	5.295,26	16,58
5	13.238,15	41,44
10	26.476,30	82,88
15	39.714,45	124,31

For example, the IMT-A RAN was assumed with throughput of 50 Mbps.

The AC algorithm was assessed in terms of the blocking probability of a call, the number of users performing inter-system handover, and the number of users decreasing their QoS in order to admit a new user and the number of users restoring their QoS (when the load of the network allows that) for different loads. Four parameters were used to define each cell: the RAN and mode type, the cell coverage, the cell location and the cell capacity. With this information every cell can be uniquely identified. All the cells have the same coverage area and capacity value. The AC and load control algorithms are based on a centralised decision.

Figure 4-7 a) and b) shows the blocking probability for the case of AC and load control without QoS renegotiation in relationship to the average throughput per cell.



a) **b)**

Figure 4-7 AC without QoS negotiation: a) Distribution of users depending on the amount of traffic per cell and b) Parameters of the algorithm.

The AC and load control algorithms perform optimally for low traffic (low average throughput/cell) and very good for rather heavy traffic (over 10 Mbps/cell). For 1.5 Mbps, the blocking probability is only 0.275 % (admission probability 99.725 %) and for 13 Mbps, the blocking probability is only 18.5 % (admission probability 81.5%). The algorithm results in a very low blocking probability when we have low or medium traffic, but although the blocking probability is still very low, it increases logarithmically with the amount of traffic. In Figure 4-7 b) the parameters of the AC are shown for the number of users (*y*-axis) and the traffic per cell (*x*-axis). The more traffic

we have in the cells, the more users are forced to handover to other networks, and the more users are blocked

The blocking probability for renegotiated QoS, just as in Figure 4-7 is estimated for different amounts of traffic per cell. Figure 4-8 shows that the blocking probability decreases when QoS is renegotiated. The blocking probability is above 10% for extremely heavy incoming new traffic per second (>10 Mbps/s). The assumed high values of incoming traffic were used to show that the algorithm performs well in extreme conditions. This in turn shows that the proposed framework is suitable for inclusion of systems with characteristics identified for next generation. For normal or medium-heavy incoming traffic, the blocking probability is around 2-3%.

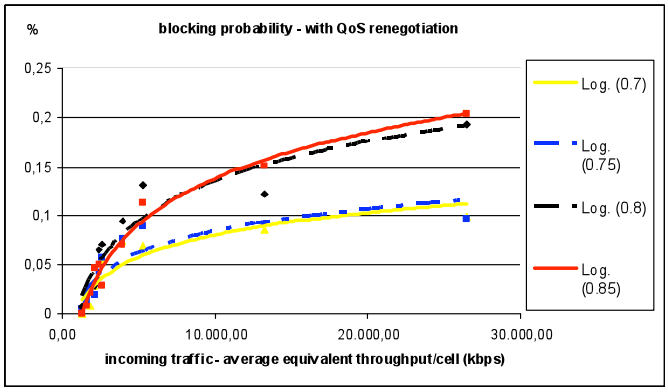


Figure 4-8 AC algorithm performance for renegotiated QoS .

In order to show the effect of the threshold values on the performance of the algorithm, four different load thresholds, 0.7, 0.75, 0.8 and 0.85 have been assumed corresponding to the percentage of used network capacity. For very high thresholds (i.e., > 0.8 or 80% use of the network capacity), however, the blocking probability is close to the estimate when no QoS renegotiation is applied. In the case of QoS renegotiation with high threshold, the QoS for almost every user that has had the QoS decreased after a very short time (almost immediately) will be restored, as opposed to the case when no QoS renegotiation is employed.

The algorithm performs very well not only for light traffic, but also for heavy traffic. With the use of QoS renegotiation, it is quite possible for new requests to be accepted to the network. Figure 4-9 shows the performance of the load control algorithm for different amounts of incoming traffic per cell.

For low traffic the percentage of the users forced to handover is low, however, this number increases as the traffic becomes heavier. The percentage of users decreasing their QoS and restoring it later also increases, but at a slow pace. This is because if the

traffic is very heavy, even the rapid decreasing of the data rate of some users will not be efficient to resolve a congestion situation. With the increase in traffic, more users will have to be dropped.

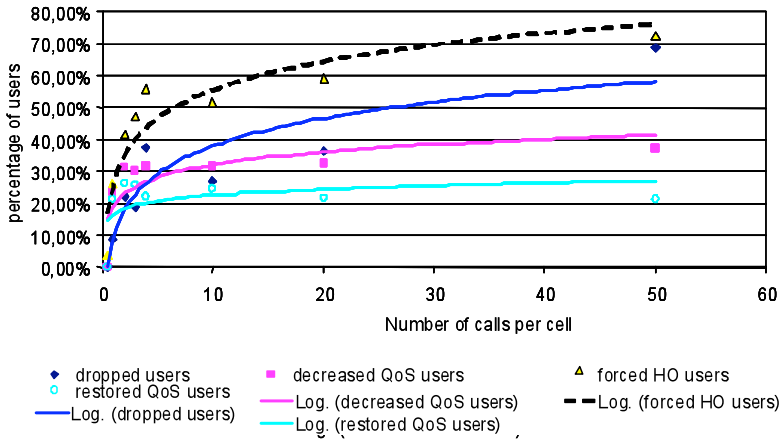


Figure 4-9 Probabilities of dropped users and users forced to handover when executing the proposed load control algorithm.

Therefore, the percentage of the users that restore their QoS (in terms of data rate) depends on the total number of users. The percentage of the users that would restore their QoS in relationship with the number of users that had their QoS decreased depends on the threshold value. The threshold for restoring the QoS is assumed 0.85. The heavier the traffic, the fewer users restore their QoS. That occurs due to the very high load of the cells.

Figure 4-10 shows the percentage of users, for which the QoS was restored. For a threshold value set at 0.7, only 6.7% of the users restore their QoS. For threshold values at 0.75 44% of the users restore their QoS, for thresholds at 0.8, 87% of users restore their QoS, and for 0.85, a 93.2% of the users restore their data rate. This means that high threshold values will allow for a larger number of users to restore their data rates.

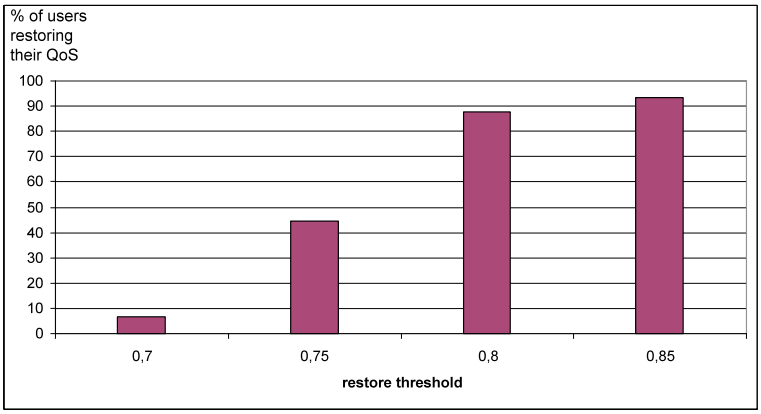


Figure 4-10 Percentage of users restoring their QoS for different threshold values.

Figure 4-11 shows the percentage of restored QoS for a fixed threshold value of the congestion at 0.85 and for different values of the average equivalent throughput per cell. Figure 4-11 shows that when the incoming traffic is heavy, a smaller number of users would restore their QoS. This is done in order to prevent that the system goes into congestion state.

For a realistic approach to QoS provision, both application and user priorities were considered. This means that a set of possible services was associated with each user profile. User profiles and application profiles can be stored in a database located outside the RAN and similar to the HIS database. The database would be beneficial also for a policy-based approach to mobility management (see Chapter 4).

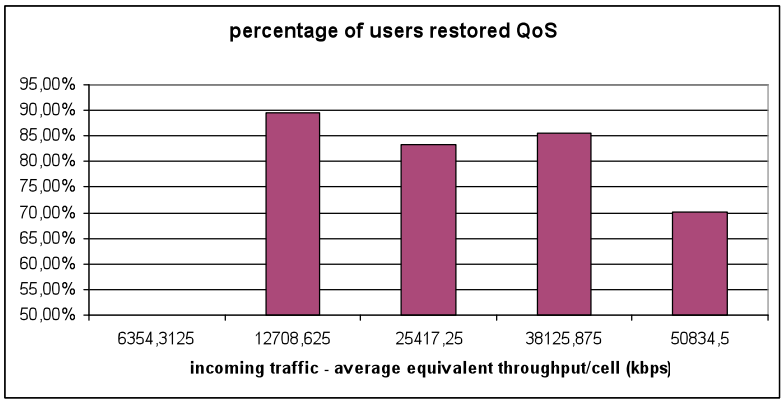


Figure 4-11 Percentage of users restoring their QoS for a congestion threshold at 0.85 and different amount of incoming traffic.

A default user profile is useful in order to minimize the probability of unnecessary handover when the user first requests an application. This type of information is referred to as static.

A complete user profile will include also dynamic information describing user activity in the network (i.e., mobility pattern, current connection at the moment of a service request). The static information includes the following:

- User ID as the identification of the user, among the different connected users;
- User subscription profile with information about the operator to which the user belongs to and the type of contract the user has. This information is used to know if the user can be granted access to the network and/or requested service. The user would be able to access a network if the network belongs either to his operator or to an operator that has agreement with the home operator. Information about the agreements between the operators is stored in the database

associated with the CoopRRM. This information is central in the process of assigning user priorities. Users can be identified based on signal characteristics mapped onto subscription profiles.

- User origin, which describes whether the user comes from a handover process initiated on another network or is a new user. A user coming from a handover session will typically have higher priority than a user requesting a new session.
- User miscellaneous information, which is meant to describe an "emergency" user. An emergency user is someone that would be highly prioritized because of his/her key role during emergencies (e.g. policemen, hospital emergency workers, firemen).

The dynamic information for a user includes the following:

- Knowledge about the RAN, to which the user is currently connected.
- User KPIs to assess the QoS the user is provided with.
- Ongoing applications to assess the resource consumption of the user.
- Application requests that are an indicator about the resources a user will need.
- User location
- User priority level.

The priority levels of the simulated users are shown in Table 4-4.

Table 4-4 User Priority levels

Contract type	User's origin	User's Basic Priority level
Emergency	Any	1
Type 1	HO	2
	New	3
Type 2	HO	4
	New	5
Type 3	HO	6
	New	7

Each application requested by a user will be associated with specific information forming a unique application profile. This information includes the following:

- Application ID is created by concatenating the user ID and the application rank of request. For example, if the user U1 requests a fifth application, this application's ID will be U1_5. In this way, for a user connection, each application will have a unique ID.

- Application service class that serves to determine the requirements in terms of rate, delay, mobility and range, and then to identify the resources needed for its operation. The service class of the application is also used to identify the compatible legacy RANs that can host the application.
- Application priority, which represents the priority level of the application, depending on its level of interactivity and rate requirement. The priorities are shown in Table 4-1. The dynamic information related to the application profiles includes the expected download time which is determined based on the duration for which we expect the user to use an application, on the type of application and of the size of the file to be transferred. Also, here the signaling and protocol overheads can significantly affect the desired download times.

The prioritization process adopted for the simulations will consider that the user priority is more important than the application priority[1]. To describe mathematically the relationship between user and application priorities, each user level is associated with a ten value, and each application level is associated with a unit value. By adding the two values, we get a total value that describes the global prioritization level of a known application used by a known user. In the case when two global levels are equal, the first arrived is first served. Global priority levels are sorted in an ordered table from high to low priority, or from the lowest global value to the highest. This priority table is updated each time a user requests a new application, during the admission control process.

In the following the cooperative congestion control algorithms are assessed for the scenario of *'busy hour'*. The goal is to evaluate the algorithms in terms of connected, blocked and dropped users. The distribution of the users for the scenario of 'busy hour' is shown in Table 4-5. Table 4-6 gives the user profiles associated to each service class.

Table 4-5 Distribution of Users during '*Busy Hour*'

Service Class	Number of users per service class	Number of users per group	Penetration factor
SC6	336	56	6, 71
SC4	18	3	0,36
SC7	846	141	16,91
SC5, SC2, SC1	60	10	1,20
SC3	1530	255	30,58
SC8	198	33	3,96
Total	**5004**	**Total**	**100,00**

[1] User priority is considered more important because it includes the user subscription profile. A user that does not belong to the IMT-A RAN will not be admitted to the network and will be handled immediately by a legacy RAN.

Table 4-6 User Profiles Associated to Service Classes

User priority	User profile	Associated service class
1	UP 1	SC5, SC2, SC1
2	UP 2	SC4, SC6, SC3,
3	UP3	SC7, SC8

The traffic generator generates a busy-hour traffic with an assumed number of incoming users about 5000 per hour. The number of connected users and their distribution according to profiles is shown in Figure 4-12.

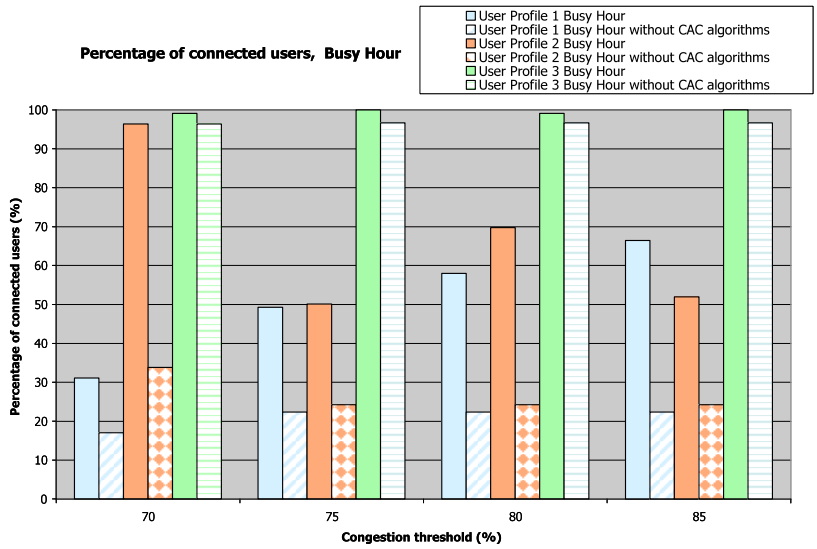


Figure 4-12 Number of connected users with and without CAC algorithms for different congestion thresholds.

The connections are given for different congestion thresholds. If the congestion thresholds are set low the number of connected users is lower and users with the highest priority are granted connection first.

Figure 4-13 gives the percentage of rejected users for each user category and congestion thresholds.

The number of rejected users of the highest priority is very low and when CAC algorithms are applied even negligible. Figure 4-14 shows the percentage of dropped users when the cooperative algorithms are applied. The highest priority users are the ones associated to less demanding service classes in terms of bandwidth (e.g., simple telephony and messaging, multimedia messaging). When the congestion threshold is set low (i.e., at 70%) the largest percentage of dropped users is in this category. This is because dropping of users is performed under action of the congestion control algorithm.

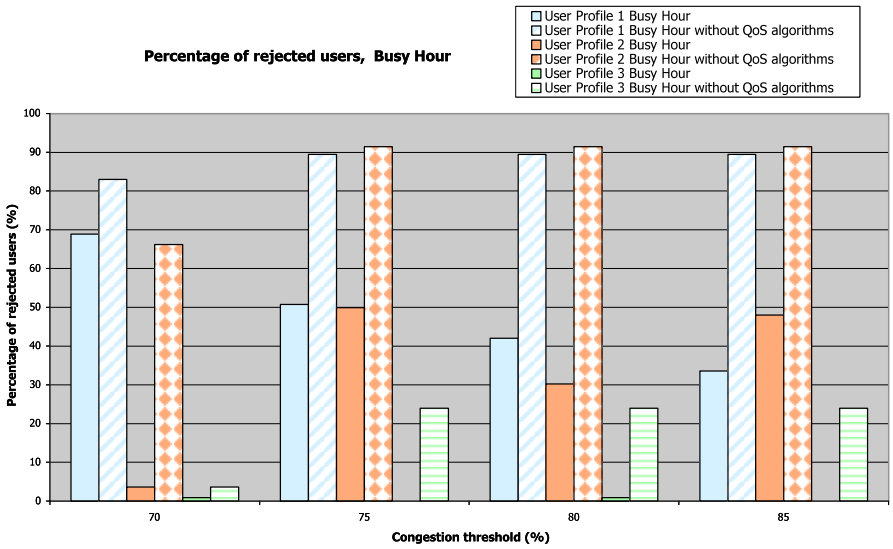


Figure 4-13 Percentage of rejected users for different user categories and congestion thresholds.

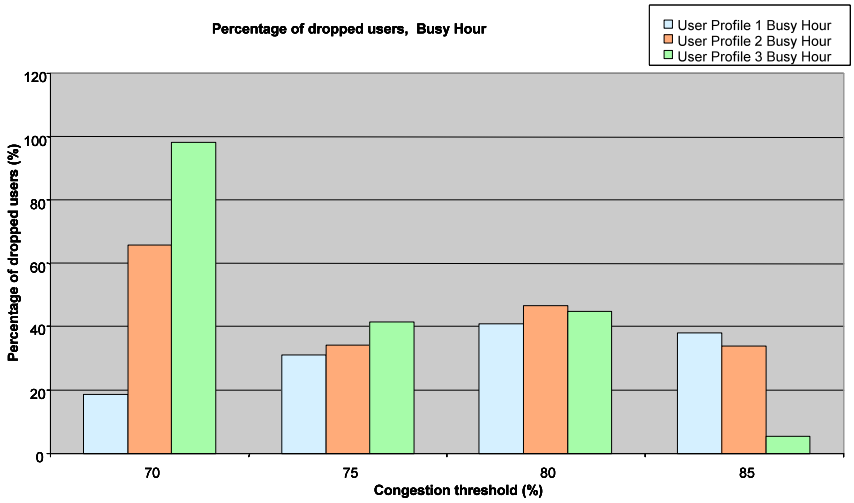


Figure 4-14 Percentage of dropped users for different user categories and congestion thresholds.

For high congestion thresholds, the percentage of dropped users belonging to the category 3 is very low but this is in accordance with the policy of operators to always reserve some capacity for simple services.

Finally, Figure 4-15 shows the results for the mean user throughput for different congestion thresholds and different traffic load scenarios.

The results have been generated to give an overall assessment of the benefits of use of cooperative congestion, admission and load control algorithms based on KPI aggregation. The mean user throughput is a KPI measured in [bps] and calculated by comparing the size of the transmitted data with the time of transmission of the data, or as given by Equation 4-27:

$$MUT_{UL/DL} = \frac{data\,payload}{time\,for\,data\,transfer} \quad \textbf{(4-27)}$$

Especially for heavy loads ('busy hour'), cooperative RRM algorithms are very important for optimised use of network capacity and QoS to users. Heavy loads would 'slow' the network down and reduce the number of handled users if QoS should be preserved.

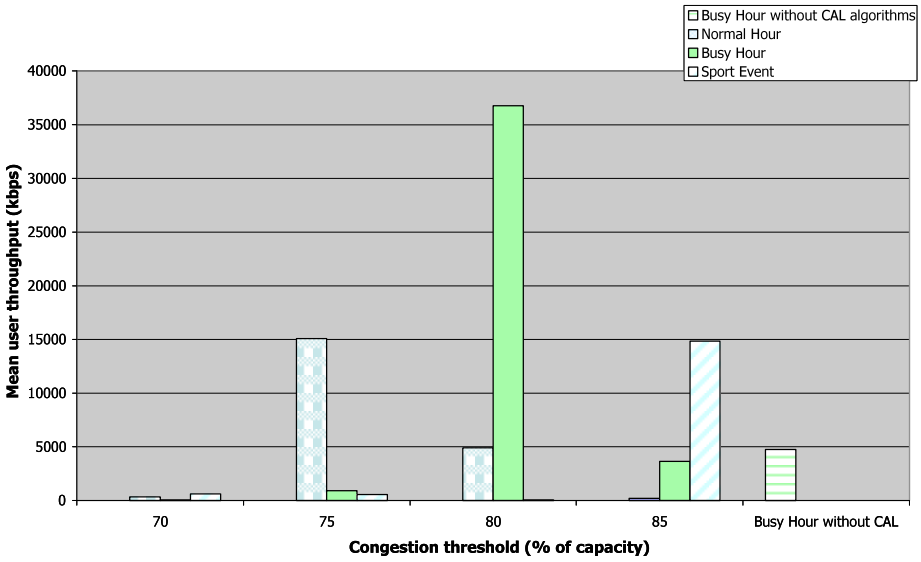


Figure 4-15 MUT for different congestion thresholds and traffic load scenarios

Monitoring of the MUT together with other KPIs, such as available bandwidth and throughput, has been proposed as a practical implementation and described in Chapter 7.

4.4 Conclusions

One requirement of the cooperation architecture is to provide some inter- and intra-RAN services such as: admission control, handover, scheduling, and QoS based management, and other services, such as billing, authentication, authorization. This Chapter proposed and assessed cooperative algorithms for admission, congestion and load control in the scope of next generation systems. Very tight coupling was selected as the for the integration of the interworking systems and cooperative RRM components. The coupling point for cooperative admission, congestion and load control is the CoopRRM or the SRRM in order to provide for an RRM framework based on the CRRM approach. This type of coupling is suitable for performing cooperative RM functions because when new and legacy systems interwork they would also have

overlapping areas of coverage. The GW pools can be beneficial for providing loose coupling possibilities in support of mobility management.

The algorithms were assessed in terms of blocking and dropping probabilities to account for the achievable loads in a given traffic load scenario.

It can be concluded that for the proposed framework load control compared to congestion control is more multi-entity related. This conclusion is used for the proposed in Chapter 5 multi-stage admission control. The simulation results justified that the proposed RRM framework introduces performance gains and scalability for catering for the future network and service requirements.

References:

[1] A. Mehrorta, *GSM System Engineering,* Artech House 1997.

[2] R. Prasad, W. Mohr, and W. Konhäuser, Third Generation Mobile Communication Systems, Artech House 2000.

[3] A. R. Prasad, and N. R. Prasad, *802.11 WLANs and IP Networking,* Artech House 2005.

[4] S., Kyriazakos and G., Karetsos, *Practical Radio Resource Management in Wireless Systems,* Norwood MA: Artech House 2004.

[5] A.-G. Acx, A. Mihovska, et al., "D1.3 Final Usage Scenarios," Deliverable 1.3, IST 2004-507581 Project WINNER, at www.ist-winner.org.

[6] Pedrycz, W., and Vasiliakos, A., *Computsational Intelligence in Telecommunication Networks, Chapter 5: Congestion Control,* CRC Press, Florida 2001.

[7] J. Perez-Romero, et al., "An Admission Control Algorithm to Manage High Bit Rate Static Users in W-CDMA," *13th IST Mobile & Wireless Communications Summit 2004,* June 2004, Lyon, France.

[8] C. Lindemann, M. Lohmann, and A. Thuemmler, "Adaptive Call Admission Control for QoS/Revenue Optimization in CDMA Cellular Networks," *Wireless Networks,* Vol.10, Issue 4, p.457-472, 2004 ISSN:1022-038.

[9] D. Gu, and J. Zhang, "A New Measurement-Based Admission Control Method for IEEE802.11 Wireless Local Area Networks," *Proc. of IEEE 2003 PIMRC*, Bejing, China, 2003.

[10] S. Black, "IEEE P802.11 Wireless LANs" Comment Resolution, March 2004.

[11] M. Frikha, et al., "Enhancing 802.11e Standard in Congested Environments," in Proc. of IEEE AICT-ICIW'06, February 2006.

[12] J. Pérez-Romero, et al., "On Managing Radio Network Congestion In UTRA-FDD,"

[13] Q. Ni, L. Romdhani, and T. Turletti, "A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN," *Wiley Journal of Wireless Communication and Mobile Computing (JWCMC),* John Wiley and Sons Ltd., 2004; Volume 4, Issue 5: 547-566.

[14] A., Mihovska, et al., "Requirements and Algorithms for Cooperation of Heterogeneous Radio Access Networks," accepted for publication in the *Springer International Journal on Wireless Personal Communications (ID WIRE 391) 2008.*

[15] E., Mino, A., Mihovska, et al., D4.4, "Impact of Cooperation Schemes between RANs—A Final Study," Deliverable 4.4 IST Project WINNER, November 2005.

[16] A. Mihovska, et al., "Algorithms for QoS Management in Heterogeneous Environments," *Proc. of WPMC'06,* San Diego, California, September 2006.

[17] P., Karamolegkos, E., Tragos, A., Mihovska, et al., "A Methodology for User Requirements Definition in the Wireless World," *Proc. of IST Mobile Summit 2006,* Mykonos, Greece, June 2006.

[18] W. Webb, *The Complete Wireless Communication Professional,* Artech House 1999.

[19] A., Mihovska, et al., "Assessment of Radio Resource Management Schemes for Efficient Cooperation of RANs," *Proc. of WPMC'05,* Aalborg, Denmark, September 2005.

Chapter 5

Policy-based Framework for Intra-System Cooperation

This Chapter proposes a policy-based framework for intra-system cooperation. The proposed framework is based on the protocols and mechanisms proposed in Chapter 2 and uses the advantages of the proposed there combined centralized and distributed approach to RRM (i.e., handover). Further, the framework uses the advantages of the data base located outside of the RAN and referred to as home subscriber server (HSS), and the one located partially in the GW (see Figure 2-1).

The goal of the proposed framework is to find an optimal trade-off between the use of centralised and distributed RRM for support of mobility management inside the RAN.

In particular the policy-based framework has been proposed in support of the following mobility management functionalities:

- RAT/BS association and selection for optimized handover control;
- User context transfer during IP and radio handover;
- Handover priority setting;
- Flow establishment and QoS class setting.

This Chapter is organised as follows. Section 5.1 defines the scenarios for policy-based mobility management. The scenarios consider interactions of mobility functions and interactions of flow handling functions related to congestion control algorithms. These scenarios serve as a basis for the proposed policy-based RRM framework. Section 5.2 proposes the strategy for RAT/BS association in order to provide for optimised intra-system handover control. Two strategies are proposed, one based on individual differentiation of the RATs/BS, and another one based on group differentiation. It is shown that the individual approach can be beneficial for introducing self-management into the BS. Section 5.3 proposes a strategy for user context transfer in two scenarios: during IP and during radio handover. The dependency of the delay and

totally transferred data is shown to depend on the polling times. Further, the strategy is based on a hybrid approach to user context transfer that involves a mandatory context transfer function which forwards only the buffered radio link control service data units (RLC SDUs) and an optional context transfer function which forwards also buffered RLC protocol data units (PDUs) in order to reduce handover delays. Section 5.4 proposes a strategy to handover priority setting and flow establishment and control. Section 5.5 concludes the Chapter.

5.1 Scenarios for Policy-Based Management

To enable efficient signaling and management between network and UT, all profiles are captured in the home operator domain for all the registered UTs [1]-[6] . The RAN is responsible for enforcement of the policy determined by the core network. The policy management is distributed between the HSS and the GW [2]. Further, mobility in the RAN is supported by traffic and control signalling from the UT to the BS that the UT is connected to, and also by the BS to BS control signalling. To ensure flexibility of the architecture, logical functionalities of the physical entities can be grouped according to the situation [1].

The scenario assumes that the interactions of the mobility management functions are as proposed in Figure 5-1. The interactions were derived from the required procedures identified for idle and active UTs in the scope of IMT-A candidate systems [1], [5]. Therefore, Figure 5-1 does not represent a complete view with all state-transitions but is rather simplified.

After power on, the UT is authenticated and authorized, a paging area update is performed (both interacting with the UT register function) and an UT micro mobility anchor is created. In idle mode only paging area updates and macro mobility functions are performed by the UT if the UT detects respective movements.

Both, the network and the UT can trigger a state change from idle to active mode. In the network this is initiated by the UT anchor point function that triggers respective paging, in the UT this is done by direct cell selection and by performing the related admission control. In this case data is exchanged with the UT register function and the UT_N micro mobility anchor.

In active mode the handover function decides about handover from one BS to another. If the decision was taken two processes run in parallel. The network performs the necessary routing changes and context transfers while the UT associates with the new cell. Finally the routing over the radio interface is updated.

This description assumes a single link between a UT and a BS. The need to support multiple links for one UT to multiple BS is for further study and is partially investigated in relation to the proposed in Chapter 4 RAT/BS association strategies.

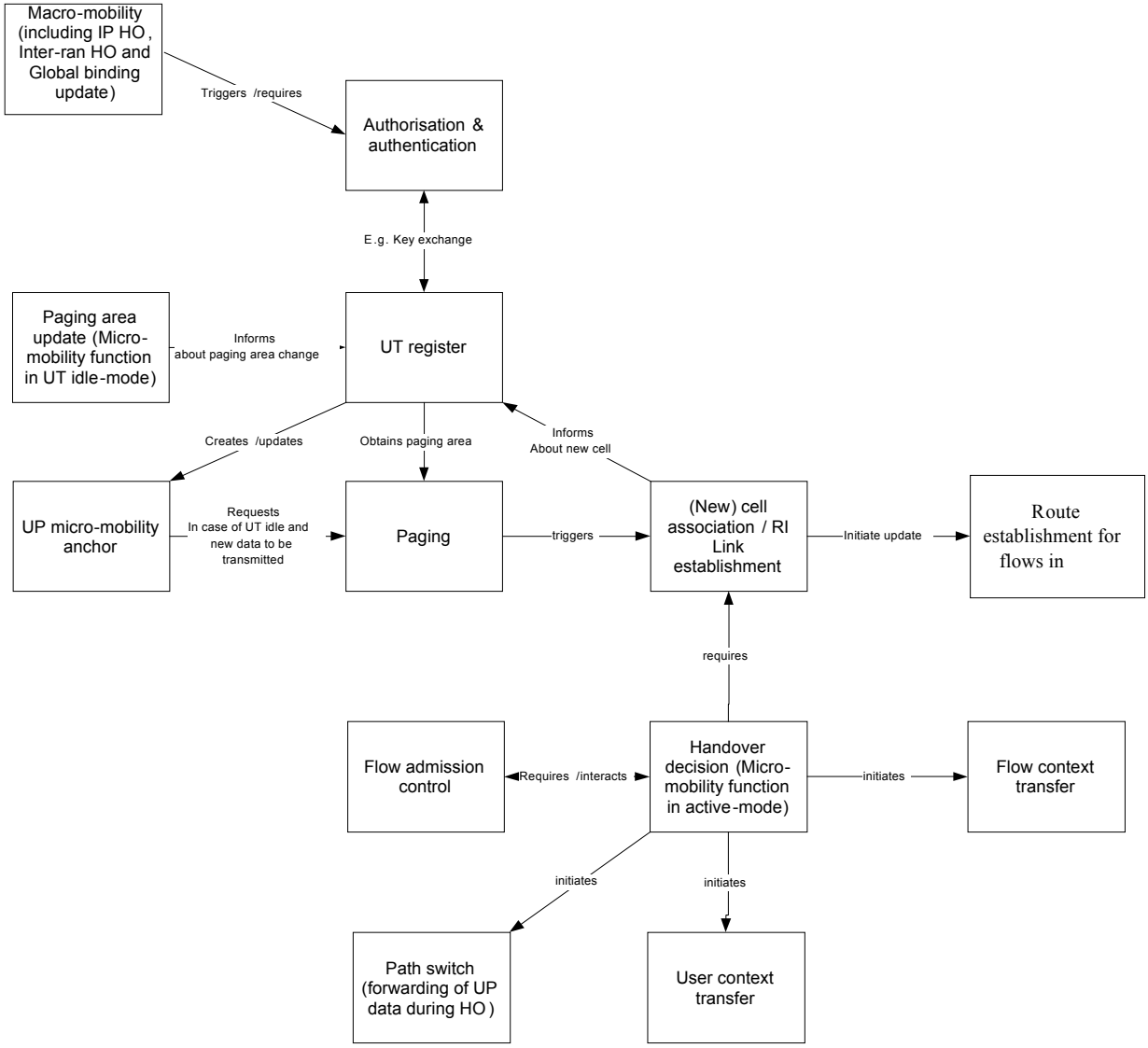


Figure 5-1 Scenario of mobility management interactions.

The procedure given here is based on the assumption that there is an UT active state where the network has detailed knowledge about the cell association and an UT idle state where only the rough location is known in order to enable power saving in the UT. The relation between the mobility functions is denoted by the text given at the arrows.

Figure 5-2 defines the congestion control interactions related to flow handling. The left part shows the flow related functions which all packets have to pass between the

ingress point of the RAN and the scheduler. The right part shows the congestion avoidance control functions and the flow establishment and release function.

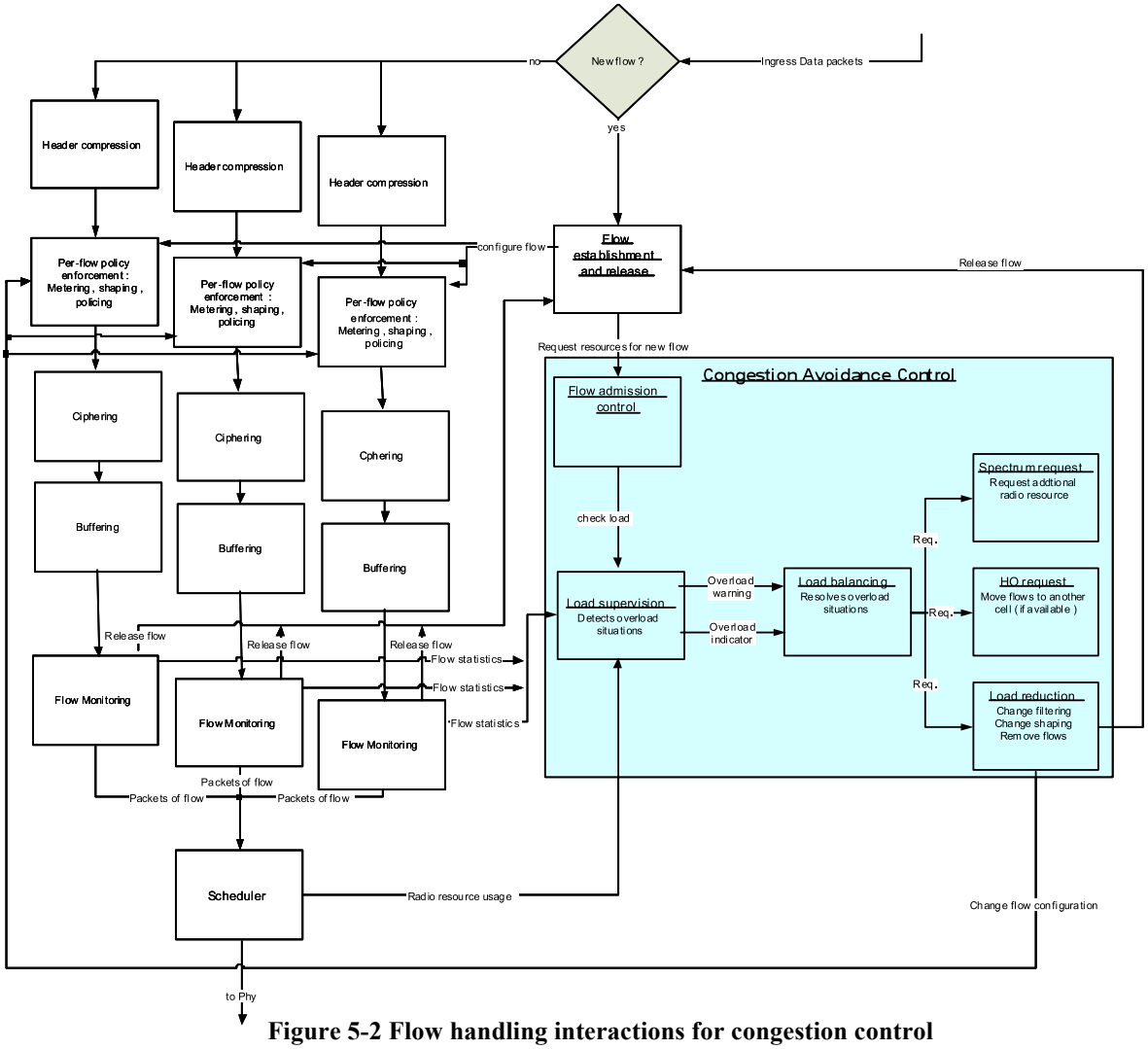


Figure 5-2 Flow handling interactions for congestion control

New incoming packets are analysed and assigned to flows. If there is no existing flow, the flow establishment function invokes flow admission control to decide on acceptance of the new flow [1], [14], [15]. If positive, the new flow will be established and header compression and per-flow policy enforcement functions are configured. This and all following packets of this flow pass the header compression and policing functions. Directly before the packets are transferred to the MAC, the packet rate over the air is measured and the activity state of the flow is detected by the flow monitoring function. After a flow has become inactive, flow monitoring triggers the release of the flow.

The load supervision gathers the flow specific load information of all monitoring functions and evaluates the load situation of the cell. After exceeding of thresholds (overload warning and overload indication), load balancing is invoked which decides on the countermeasures to resolve the overload situation. The requesting of handover of

flows to other cells or traffic reassignment to another BS as a means for QoS handling are assumed for the proposed policy-based RRM framework. If none of these exist, load has to be reduced, by changing the QoS policing parameters up to dropping all packets of a flow.

A protocol reference architecture is proposed to comply with interactions defined by the two scenarios [14], [15]. This architecture is shown in Figure 5-3 a) for the user plane interactions, and b) for the control plane interactions.

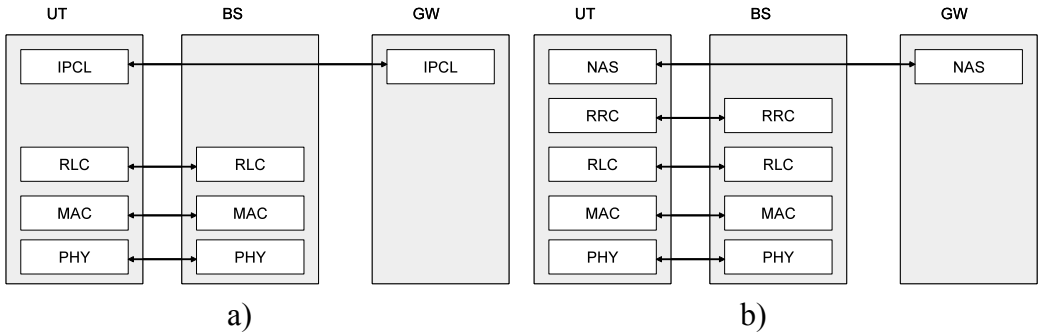


Figure 5-3 Protocol reference architecture for the proposed scenarios.

In Figure 5-3 a) in the case of transfer of user data, the IP convergence layer (IPCL) adapts the higher-layers data flows (e.g., IP packets) to the transmission modes of the radio link control (RLC) layer, establishing the transfer data protocol with a peer IPCL entity, compressing the long IP headers and ciphering the IP payload. A user plane connection (i.e., an IPCL layer session) can generate several RLC layer flows using different QoS classes (see

Figure 5-2). The IPCL layer protocol performs a two-fold function: it takes care of the transfer of user plane data between two IPCL layers in different nodes (e.g. in the UT and the GW) and of the IPCL services for handover [1], [2]. Transmission of user data means that the IPCL receives an IPCL SDU and forwards it to the RLC layer and vice versa. In this way TCP [7] /UDP /IP packets [8] are transmitted as IPCL PDU packets (composed of an IPCL header and an IPCL SDU). An RLC entity receives/delivers RLC SDUs from/to upper layer and sends/receives RLC PDUs to/from its peer RLC entity via lower layers. The problem arising from this architecture scenario is that because the RLC protocol terminates in the BS, the *automatic repeat request* (ARQ) mechanism normally employed to support TCP performance is not sufficient to support lossless mobility support [9], [10], [11], [12], [13].

In the control plane [Figure 5-3 b)], there is a Non Access Stratum (NAS) protocol over the radio access protocols terminating at the GW. The NAS control protocol is important in reference to a number of functions related to radio access (e.g., paging) [1].

The radio resource control (RRC) layer handles the controlling functions and signalling to one dedicated UT (in RRC active mode), paging of idle UTs and broadcasting of system information. Most of the RRM functions proposed in Chapter 2 are performed based on the RRC messages.

5.2 Policy for Handover Control during RAT/BS Association

The objective of the proposed policy is to provide for handover control during RAT association in an indoor scenario. It is based on the defined mobility scenario interactions in Figure 5-1. A hierarchical control structure was proposed in Chapter 3 in compliance with the requirements identified for next generation systems, which assumes that BS_{WA} would overlap the coverage of the BS_{LA}. This means that UTs on the border of the coverage of the BS_{LA} would also connect to the BS_{WA} and maintain simultaneous links. This leads to unnecessary resource use. The proposed here strategy differentiates UTs during their initial association according to the probability of performing a handover to the BS_{WA} and connects those UTs directly to the BS_{WA}. The scenario is shown in Figure 5-4.

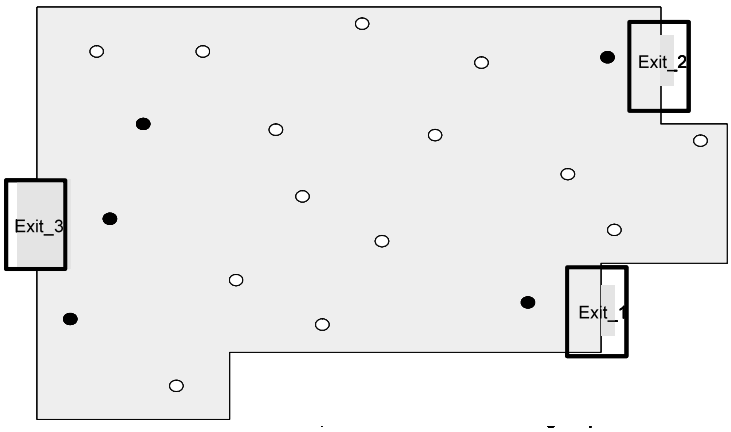


Figure 5-4 Coverage area of BS_{LA} including border BS_{LA}.

In the assumed scenario, several exits are available and the BSs are spread almost uniformly inside the building. Figure 5-4 shows the BSs, from which the UTs may potentially lose their LA coverage as dark circles. These BSs are referred to as border BSs.

In order to differentiate the BSs_{LA}, several metrics can be used. Here the following assumptions are made:

- A central entity is in charge of the handover;
- The handover area is handled by the strategies discussed earlier and the handover procedure itself is already defined.

The applied metrics are group differentiation and individual differentiation.

5.2.1 Group Differentiation

In this case, for each BS_{LAi}, $i = 1,.., N$, the probability P_i is defined as the probability that UTs, associated to the BS_{LAi} would perform a handover from the LA to the WA domain. P_i is updated each time a handover is completed. The sum of the probability over all the BSs is equal to 1 or

$$\sum_{i=1}^{N} P_i = 1 \qquad \textbf{(5-1)}$$

Let k be the k^{th} handover from a BS_{LA} to a BS_{WA}. Then for each BS_i the probability for performing a handover, is defined as:

$$P_i^{(k+1)} = P_i^{(k)} \times \frac{k}{k+1} + \frac{A_i^{(k)}}{k+1} \qquad \textbf{for } k \geq 1, \ \forall i = 1,...,N \qquad \textbf{(5-2);}$$

where

$$A_i^{(k)} = \begin{cases} 1 & \text{if the UT was associated to } BS_i \\ 0 & \text{otherwise.} \end{cases}$$

Here, two policies are introduced for the initialization of the error probability. The first policy intializes all probabilities, for $k = 1$, to the same value:

$$P_i^{(1)} = \frac{1}{N}, \qquad \forall i = 1,...,N \qquad \textbf{(5-3)}$$

If the central entity that is in charge of the handover (i.e., RRM server) is aware of the LA network topology and the locations of the BS, the probabilities for $k = 1$ may be initialized based on this information. For example, let Ω denote the set of border BSs_{LA}. In Figure 5-4, the cardinal of the set Ω is equal to 5. In this case, a relevant initialization is as given by Equation 5-4:

$$P_i^{(1)} = \begin{cases} \dfrac{1}{Card(\Omega)} - \varepsilon & \text{if } i \in \Omega \\ \varepsilon & \text{otherwise} \end{cases} \qquad \textbf{with } \varepsilon \geq O \qquad \textbf{(5-4)}$$

Convergence to stable probabilities is quite rapid for both types of initialization. Sometimes, it may be useful to reinitialize the probabilities during the process, after *p* handovers. The value of *p* may be configurable.

Finally, each *Q* handovers or each *T* seconds (or minutes), the state of each access point is updated according to:

- If $P_i > \gamma$, then BS_i is declared as border BS;
- Otherwise, BS_i is set to non-border BS; where γ is a threshold, typically of 10^{-3}.

Each UT associated to a border BS is required by the RRM server entity to set up two active links, on the LA and the WA networks, respectively. This means that each UT, associated to a border BS, has to be authenticated and registered on both networks. On the other hand, if an UT is attached to a non border BS, then it may release its WA link, as it will probably not be handed over to the WA domain. The BS states and associated probabilities are stored in a database at the RRM server entity. Consequently, the activation or not of the WA link may be ordered by the RRM server, as soon as a change in the BS states is detected, instead of considering a periodic update.

If a BSs_{LA} is added, removed or even moved in the LA domain, after a few iterations its probability will rapidly converge to a stable state, which will reflect its new location. Such dynamic update is a kind of self-organization of the BSs_{LA}. Moreover, the database containing the probability associated to each BS may be used to monitor the LA network. After a few iterations, if the network is stable (e.g., BSs are not moved or removed), it is possible to draw a map with the BSs location and check if this map is coherent or not. High probabilities may be used to detect abnormal handovers due for example to a malfunctioning of a BS.

5.2.2 Individual Differentiation

In the *group differentiation* metrics, the sum of probability over all the BSs is equal to 1. The applied condition for executing the policy, rated the distance of each BS_{LA} to the WA domain, in relation to each other. In other words, when a handover towards BS_{WA} occurs for one specific BS, it affects the probability related to all other BS_{LA}.

An alternative is to define an absolute probability associated to every BS_{LA}, characterizing the probability that users associated to it will require a handover to the BS_{WA}. For a given BS_i, four types of events can occur:

- Incoming session initiated within the BS_{LA} coverage;

- Incoming session resulting from a handover (session already existing outside the BS_{LA} coverage). The latter could be subdivided into: *(i)* intra LA; *(ii)* handover from WA. The number of these events is denoted as A_i;
- Terminating session within BS_{LA} coverage;
- Terminating session due to handover. Here, again, two events can be distinguished: *(i)* intra LA handover and *(ii)* handover to WA. The occurrences of these events are denoted as B_i.

From the above, Equation 5-5 defines the absolute probability for the occurrence of a handover as:

$$P_i = \frac{A_i + B_i}{E_i}, \qquad \textbf{(5-5);}$$

where E_i denotes the total number of events, as defined above, for the BS_i. P_i then characterizes the "distance" of BS_{LAi} to the BS_{WA}.

Without any *a priori* information about the BSs location, the initial probability could be set as 0.5 (no specific preference). This probability is then used to determine whether or not the terminal should keep/initiate its WA session alive, when associating with a specific BS.

Other metrics to differentiate the BSs are possible. The idea is to use any suitable metrics to differentiate the BSs in order to reflect that handover to/from a BS_{WA} does not occur with the same probability for the different BSs_{LA}. Such method is suitable for a system without strong central control and O&M interworking [5] . The heuristic probability updates provide awareness of the handover tendencies, thus realising a *self-learning* RAT association mechanism.

5.2.3 Advanced Function for Handover Optimisation

It is further proposed to implement a similar method in the RRMserver entity as an advanced function for the handover process. The method would optimise the trigger for the handover (see Figure 2-7). The proposed method, based on a "relative" location of the BS_{LA} to the BS_{WA}, can limit the useless handover preparations and conversely anticipate handover when it is most likely required.

The following shows the effect of a network performance trigger that combines the measured SINR and the network load. The metric used to activate the trigger is the *residual throughput* and is defined as:

$$\textit{Data Rate} * (1 - \textit{PER}) * (1 - \textit{Channel Occupation}) \qquad \textbf{(5-6)}$$

The word "residual" means in this context, that if a part of network resource is already occupied by other users, then the handover decision is only based on the remaining available for this UT bandwidth. The handover trigger is based on a comparison between the estimation of the *residual throughput* on the current cell (i.e., *current_residual_throughput*) with the one that could be achieved on another cell (i.e., *target_residual_throughput*).

If the ratio between *target_residual_throughput* and *current_residual_throughput* is bigger than the following pre-defined threshold:

target_residual_throughput/current_residual_throughput > throughput_threshold

then the handover is triggered.

To derive the *Data Rate,* the *Channel Occupation* and the *PER*, the UT performs measurements in the used cells and on broadcast messages sent by the BSs of neighbouring cells (see Chapter 2). To show the effect of this trigger, the above scenario is modified to involve three UTs; UT1 and UT2 are initially in the coverage overlap area of the BS_{WA} and later move towards the center of the area covered by the BS_{LA}; UT3 is always located inside the building (coverage of BS_{LA}). Figure 5-5 shows the triggering of the handover based on the measured by UT2 throughput on the physical link and without setting a threshold (i.e., threshold → ∞).

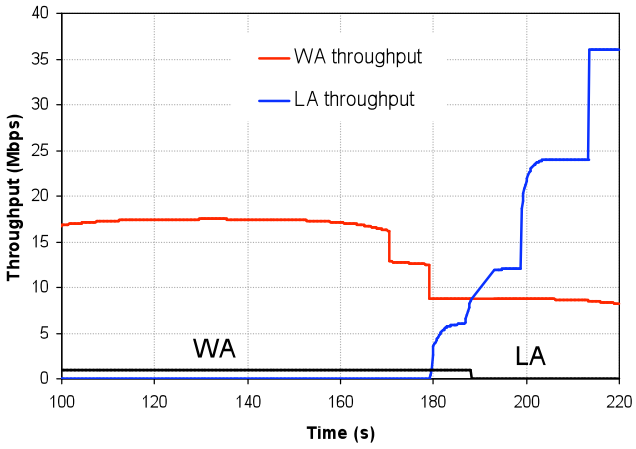


Figure 5-5 Handover triggered by measured throughput

The UT initially connected to the BS_{WA}, continues to use it until 190s, and then switches to the BS_{LA}. Figure 5-5 shows how the throughput estimate in the WA reduces as the UT moves away from the BS_{WA}. At 180s the UT enters the coverage of the BS_{LA} and starts to estimate the throughput on that mode. The LA throughput increases as the

UT approaches the BS_{LA} and at 190s it outcomes the WA throughput and thus the handover from WA mode to LA mode is performed.

Figure 5-6 shows the results for a threshold based on the measured network load (i.e., bandwidth occupation) as the advanced metric to trigger handover.

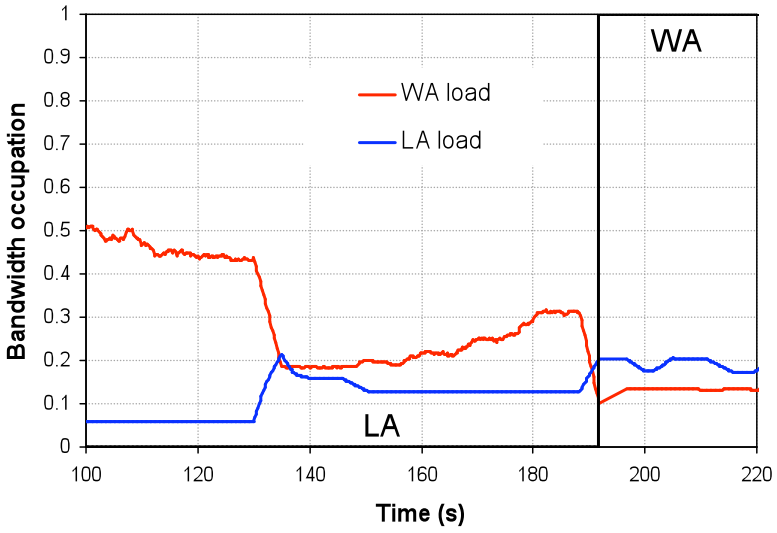


Figure 5-6 Effect of bandwidth occupation as a trigger of handover.

The bandwidth occupation on the WA decreases when the UTs handover to the LA mode (i.e., the UT3 performs handover at 135s while the UT2 at 190s). It is obvious that such an approach in fact balances the network load.

It can be concluded that use of network load as a trigger is better from network capacity optimization point of view, whereas, the throughput measure is more efficient for the individual UT differentiation.

Combined use of triggers and policy strategies can enhance the handover process, while keeping the process transparent to the user. The above strategy can also be applied in the context of inter-system handover between a WA system and a short-range system when the additional delays can be expected from the processes of registration, authentication, authorization and so forth. For the inter-mode scenario security and registration procedure may be less stringent as one can imagine that a single operator controls the various types of deployment. However, a scenario when several operators are active in the LA deployment, the security procedures must be applied.

5.3 Network-Controlled Flow Control for User Context Transfer

In a flat architecture [4], [5], an intra-system handover (between BSs) of a UT takes place by switching a tunnel in an anchor point in an anchor node (i.e., GW). The context of a UT for the architecture shown in Figure 2-1 is established mostly in the GW after

successful initial access [4]. It is conceptually separated from the flow contexts of this UT in order to enable individual routing of these flows through different cells [2]. The user context might be kept centrally in the RAN or moved from BS to BS in the case of a handover. The user context contains all information related to a UT in the system: user-id; flows associated to the user; physical context information, (e.g., location, direction of movement, speed).

5.3.1 Radio and IP Handover

The context of a flow is established and released by the flow establishment/release function (see

Figure 5-2). If a flow should be moved from one BS to another, the context of the flow has to be transferred to the new BS. This might happen in the case of *radio handover* due to a changed radio link condition or if multiple radio links are available to the UT (overlay cells) due to the shift of load between the cells.

The transfer could happen within a one domain (micro-mobility) or between two domains (macro-mobility). In the case of macro-mobility, IP handover is required. Chapter 3 introduced the concept of pool of GWs (see Figure 3-8 and Figure 3-9) made possible by the proposed cooperative RRM framework. This concept minimizes the need for an IP handover (i.e., macro-mobility). Therefore, the GW association will be preserved even when a UT performs a handover to a BS that is controlled by a different GW. In that case, a change of the IP address is not necessary. To preserve the scalability of the GWs, a shift in the user context is necessary from the highly loaded GWs to the less loaded GWs. This is performed by the load balancing algorithm, (introduced in Chapter 4 and further enhanced in Chapter 6). In that case two handovers will be performed; an IP and a radio handover (i.e. hybrid handover). The required minimum signaling during hybrid handover is shown in Figure 5-7.

The IP handover will be performed when the user changes GWs and this would result in a change of the IP address. Radio handover also is performed when the user changes the BS to which the UT is connected and a new connection to a BS that is controlled by the other GW is established.

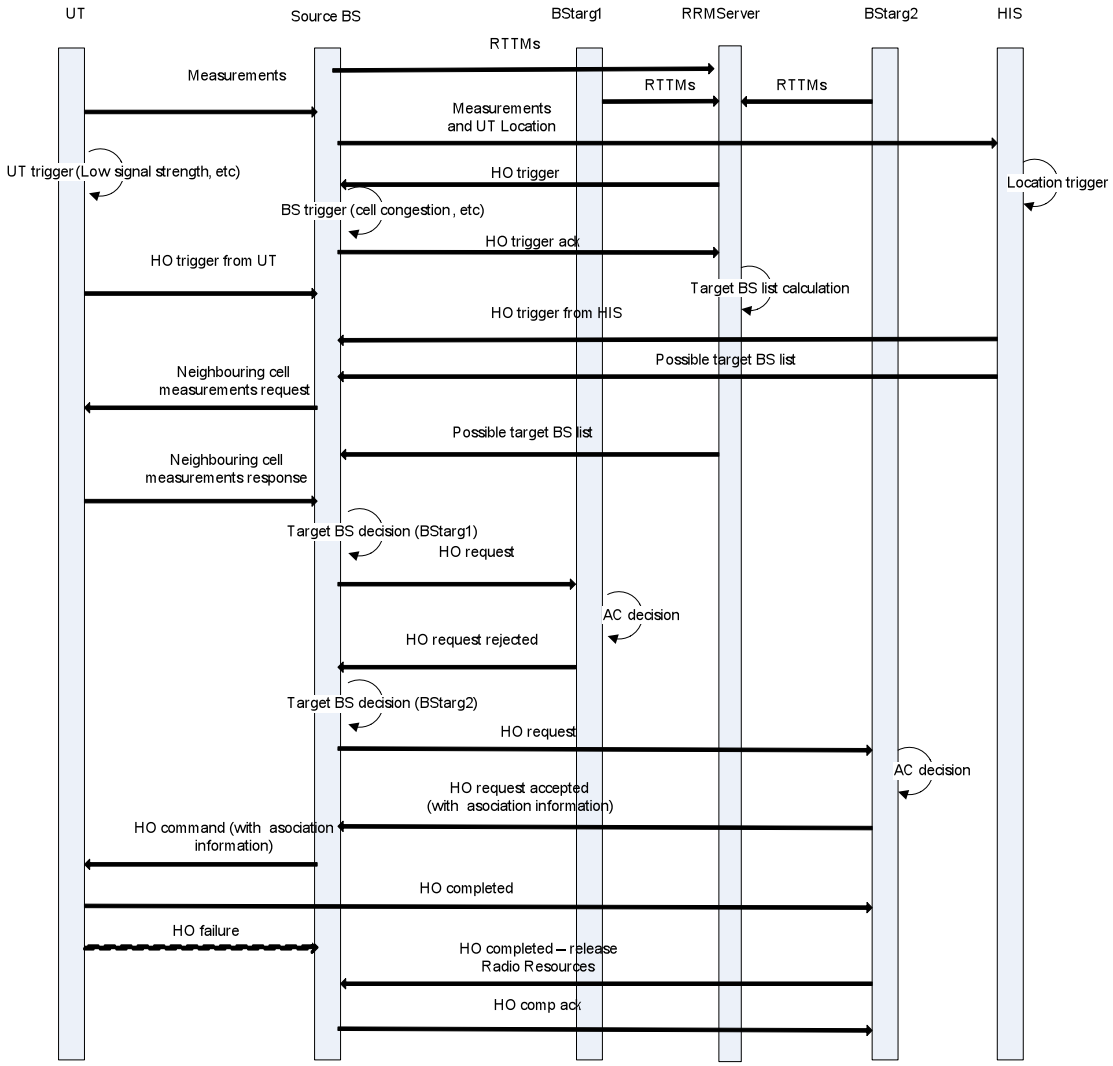


Figure 5-7 Required signaling during hybrid handover.

The context transfer of control information is needed to make the handover decisions, and to avoid re-establishment and re-authentication of connections.

The context transfer of user plane information is needed because there is no flow control similar to the functionality of the I_{ub} interface in UTRAN [16]-[18] envisioned for the RAN of IMT-A candidate systems [4], [5] where fast connections must be maintained throughout the system, which results in large amount of data being buffered at the BS. RLC SDU context transfer was proposed for the LTE system [19], [20].

It is proposed here based on the protocol reference architecture in Figure 5-3 that user context transfer during radio handover is performed by forwarding of user data in the following two ways:

- Forwarding of buffered RLC SDUs as a required function;
- Forwarding of buffered RLC PDUs as an additional function in order to decrease delays due to retransmissions that have occurred during handover.

The forwarding process is further assisted by a proposed policy-based flow control mechanism that serves to avoid overflows of buffered data at the BS and GW.

5.3.2 Message Exchange for Policy-Based Flow Control

Due to the capacity difference between the GW and BS, overflow of user plane data at the BS might happen for a single UT or a group of UTs. It is proposed here that if the BS detects that the buffered data of a particular flow is rapidly increasing and approaching the buffer limit, it would issue an explicit signal to the GW to suspend the forwarding to the BS, in addition to preventive buffer management policies such as random early dropping (RED) [21], [22]. This explicit signalling conveys two fields, one is necessarily the identifier of the IPCL data flow of the UT (see Figure 5-3) that is approaching the buffering limit, and the other is a command that requests the GW to suspend or to resume the data forwarding to the BS. Such a command is identified by the IPCL layer in the GW, so it would temporarily hold its PDUs in the buffer, instead of forwarding them immediately to the RLC in the BS after the processing of SDUs. If later the congestion state for the flow at the BS is alleviated, the BS could again send a recovery message that informs the GW to increase the forwarding rate.

A possible solution is to use GW-BS flow control signalling. The GW-BS flow control signalling during handover is implemented with the *XON/XOFF* flow control [23]. This proposed implementation is shown in Figure 5-8.

It is assumed that at the start of the handover at least one downlink (DL) packet flow is in process prior to the handover, and that the QoS requirements of the UT for on-going packet flows are known at the GW.

During the handover preparation, the sequence number of the last packet in each QoS flow assumed to be successfully transmitted to the UT is also sent to the source BS. The BS-GW signalling is utilized to help maintain the in-sequence delivery of IPCL PDUs from the GW to the BS. The source BS also has to notify the GW about the switching of the data forwarding path after the handover to the target BS is completed. Before the notification of the path switching, the GW would still forward the IPCL PDUs to the source BS, which then needs to be tunnelled by the source BS to the target BS. If the forwarding of the data is not finished before the path switching, new IPCL PDUs from the GW might arrive at the target BS before some of the previous IPCL PDUs from the source BS. The RLC layer on the target BS cannot recover the original

order of the IPCL PDUs, unless it could probe into the IPCL header to obtain the sequence number information, which violates the layering paradigm.

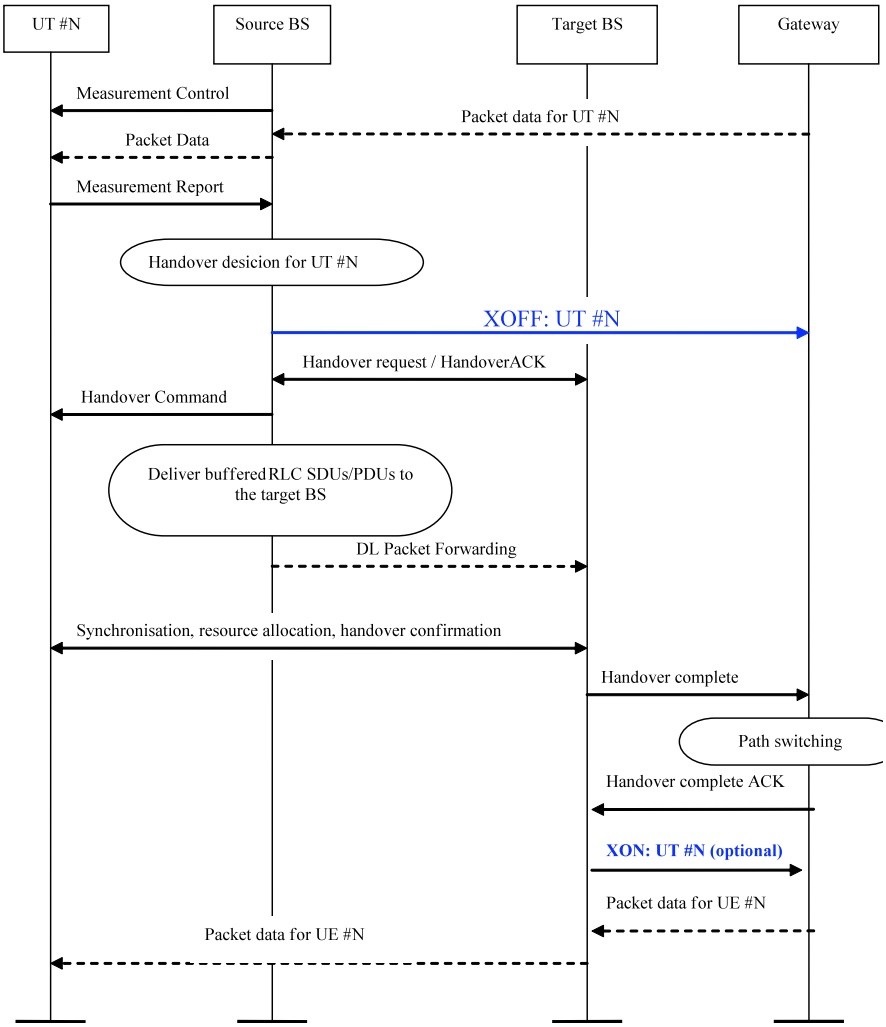


Figure 5-8 GW-BS Flow Control Signalling during Handover.

The source BS could request the GW to stop forwarding further IPCL data to it, to avoid unnecessary tunnelling of the PDUs to the target BS. In addition, the source BS should notify the GW about the path switching after it has forwarded all the buffered RLC SDUs/PDUs to the target BS. As a result, the forwarded data from the source BS always comes earlier than further IPCL data from the GW to the target BS, so that in-sequence delivery could be preserved. In the whole handover process, the network policy classifies users that allow higher QoS tolerance, and, therefore, may support the flow control.

In the proposed here XON/XOFF flow control mechanism the receiver distinguishes only two different states: *'ready'* and *'not ready'* to accept data. *XON/XOFF* represents the command to suspend/resume data forwarding from the GW

to the BS. The transmitter, upon acquiring a *'ready'* signal, transmits data at an arbitrary rate until it acquires a *'not-ready'* signal.

After that the transmitter does not transmit any data packets until a *'ready'* signal is again acquired.

The delays associated with the proposed signaling is shown in Figure4-9.

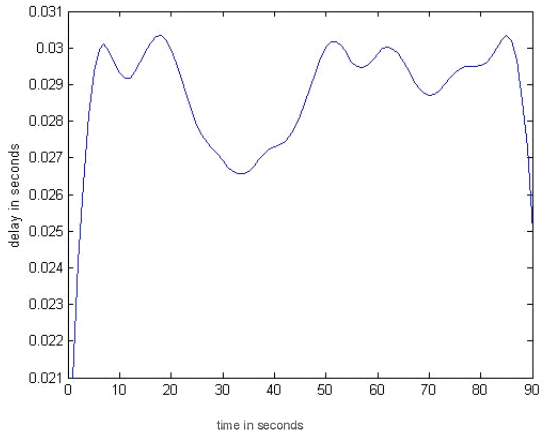


Figure 5-9 Delays associated with user context transfer signaling.

Another possibility is to transmit user data in another type of frame than the frames required for executing flow control. This means that flow control can be based on out-of-band signalling (similar to the use in HDLC protocols). The receiver then will issue *'receive-not-ready'* frame when all its buffers are full. As soon as it can accept new data, the receiver sends a 'receive-ready' frame. This frame contains also the sequence number of the next expected data frame. A high-level goal should be to minimize the amount of out-of-band signalling and should preferably be restricted to information that is beneficial for the receiver to know before decoding the received packet.

Instead of having a simple flag to stop/resume the GW-BS data forwarding, the BS could also advertise a receiving window size to the GW, so that the IPCL layer on the GW would only forward as much data as allowed by the receiving window of the BS. Such an advertisement of receiving window by the BS to the GW could be updated periodically or be event driven, (e.g., in case of handover or other situations that suddenly decrease the air interface data rate for a particular UT).

The flow control procedure can be further enhanced by introducing flow control policies based on the QoS requirements of each flow that is buffered at the GW prior to handover. Flows can be distinguished as *time-critical* (i.e., for real-time applications of the *conversational* or *streaming* classes) or *non-time critical* (i.e., associated to a

background or interactive service class, e.g., download of a file, email). Further, flows can be *critical,* which means that such a flow would require secure delivery (i.e., a low drop probability and relatively slow context transfer would be expected). For each flow a specific buffering policy would be executed at the GW.

This topic is envisioned for further work.

5.3.3 Policy-Based Forwarding of RLC SDU and RLC PDU

This Section investigates the effect of RLC context transfer of RLC PDUs, in addition to RLC SDUs, for reducing the delays due to handover. Further, it proposes strategies for the two types of proposed forwarding that comply with the specifics of the proposed RRM architecture.

5.3.3.1 Policy for RLC SDU Context Transfer

During RLC SDU context transfer, all data stored in the SDU buffer of BS_{curr} is transferred to the SDU buffer of the $BS_{targ.}$ No particular processing of the SDUs is required. The SDUs are embedded in context datagrams and then are transferred to the $BS_{targ.}$ This process is shown in Figure 5-10.

The following policy is proposed for forwarding of outstanding data in the buffers of the old BS.

In the DL:

- Before handover, the RLC receiver, in the UT, transmits a status message to the source BS.
- During handover, the RLC sender, in the source BS, forwards all buffered RLC SDUs to the target BS.
- After handover, the target BS transmits all RLC SDUs that were forwarded from the source BS to the UT.

In the uplink (UL):

- Before handover, the RLC receiver in the source BS forwards all successfully received RLC SDUs to IPCL in the GW.
- After handover, the UT transmits RLC SDUs.

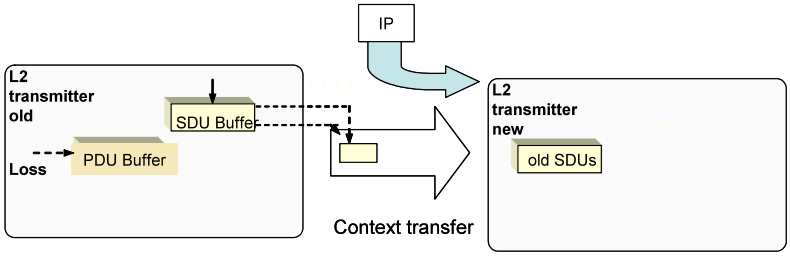


Figure 5-10 Forwarding of RLC SDUs.

In order to avoid inconsistent states between sender and receiver before handover, it is proposed here that in the DL, the RLC receiver in the UT is instructed from higher layers to transmit a status message to the source BS before handover. If the status message is successfully received, some RLC SDUs may get acknowledged and less RLC SDUs may need to be forwarded. If the status message is lost, RLC SDUs that have already been successfully received may unnecessarily be forwarded and retransmitted. The status message can be formulated as shown in Figure 5-11.

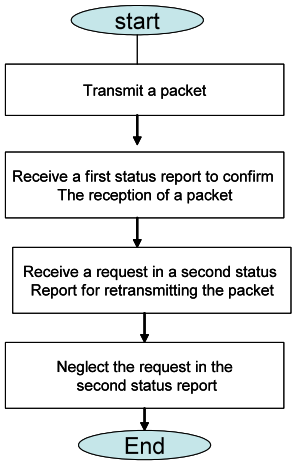


Figure 5-11 Structure of a status message.

Data forwarding, such as context transfer of RLC SDUs, increases performance significantly [24]-[26]. The gain from data forwarding is higher when a lot of data is buffered at the BS. Bandwidth, error rate and round trip time have an impact on the pipe capacity. If the pipe capacity increases, then more data is buffered in the source BS and, hence, more data are lost if buffers are discarded instead of forwarded during handover [26].

The extra delay caused by the unnecessary retransmission of RLC SDUs may have a negative impact on the performance of higher layer protocols. New data is delayed, if

the buffered RLC SDUs are retransmitted first. Therefore, it is proposed that RLC PDUs are also forwarded in some cases.

5.3.3.2Policy for RLC SDU and RLC PDU context transfer

When both RLC SDUs and RLC PDUs are forwarded, the RLC SDUs are divided in smaller sizes of RLC PDUs. This is shown in Figure 5-12.

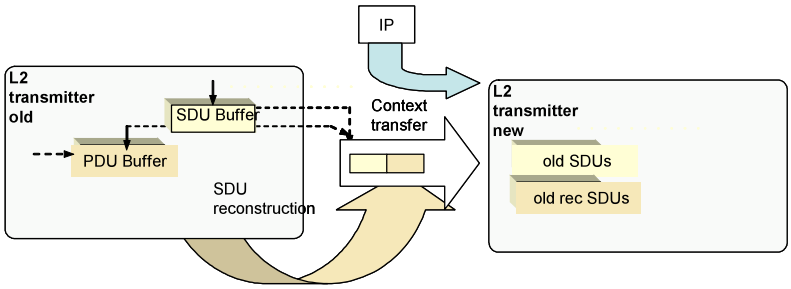


Figure 5-12 RLC SDU division in RLC PDUs.

With RLC PDU context transfer, if RLC SDUs are divided into many RLC PDUs of smaller size and only a few of the RLC PDUs that belong to one and the same RLC SDU are lost, then it is not necessary to retransmit the whole RLC SDU (as it is if only context transfer of RLC SDUs is performed).

The following policy is proposed for the forwarding of RLC PDUs.

In DL:

- Before handover, the RLC receiver in the UT transmits a status message to the source BS.
- During handover, the RLC sender in the source BS forwards buffered RLC SDUs and RLC PDUs in the transmission queue to the target BS. Also the RLC state is transferred.
- After handover, the UT transmits a status message to the target BS, and the forwarded RLC PDUs that have not yet successfully received by the UT are transmitted.

In UL:

- Before handover, the RLC receiver in the source BS transmits all successfully received RLC SDUs to IPCL in the GW and any remaining RLC PDUs to the target BS. Also RLC state is transferred.

- After handover, the target BS transmits a status message to the UT. The UT continues to transmit RLC SDUs and unacknowledged RLC PDUs.

Just as for RLC SDU context transfer, inconsistent states should be avoided if possible. Therefore, it is proposed that, in the DL, the RLC receiver in the UT is instructed from higher layers to transmit a status message to the source BS before handover.

In the UL, this seems less important, since the RLC sender remains in the same node, the UT, after handover. It is further proposed that higher layers instruct the RLC receivers to transmit a status message after handover. If the sender fails to get status before handover, unnecessary retransmission could be avoided with a status message after handover.

5.3.3.3 Efficiency of the Proposed Policies

RLC PDU context transfer may also be efficient even if there is a one-to-one mapping between RLC SDUs and RLC PDUs. The benefit is explained as follows.

If assumed that the receiver has successfully received a number of RLC PDUs before handover, but the sender has not received any status information, which could occur if a poll request or status message is lost, or if the poll interval is long. If only RLC SDU context transfer is applied, then the received but unacknowledged RLC SDUs need to be retransmitted after handover even though it is unnecessary, since they were successfully received already before handover. With RLC PDU context transfer, on the other hand, the sender could poll the receiver after handover and get status information about the successfully received data. Thus, unnecessary retransmissions could be avoided.

5.3.3.3.1 Polling Requests and Status Messages

One of the more important questions regarding the efficiency of RLC is how often poll requests and status messages should be exchanged. Too frequent transmissions waste radio resources, increase power usage, and may even trigger unnecessary retransmissions [27]. Too infrequent transmissions may instead stall the RLC transmission window, which may result in under utilisation of the radio link. In 3GPP RLC [20], a multitude of options for polling are specified, some of which are one-shot (when some condition becomes true, e.g., last PDU in buffer) and others which are recurrent (expiry of poll timer and periodic polling). Incorrect configuration may cause

deadlock. Recurrent polling is required to avoid deadlock [28]. In order to avoid too frequent transmissions, a *poll prohibit* timer and a *status report prohibit* timer can be used. A *poll prohibit* timer sets the limit for the minimum interval that is allowed between two poll requests, and a *status report prohibit* timer between two status reports.

Here a polling with a poll timer is proposed. Then the amount of outstanding data that is buffered in the source BS at handover depends on the available bandwidth, the delay and the setting of the poll timer, $b(T + d + d)$ where b is the available bandwidth, d is the one way delay over the radio link, and T is the timeout value of the poll timer. The amount of outstanding data is the product between the bandwidth and the time required to transmit the data, which is the timeout value, T, and the time it takes to transmit a poll request and to get a status message back, $d + d$. Data are assumed to be transmitted also while the poll request and status report are exchanged. When the RLC sender receives the status report, it determines if retransmissions are needed or if new data can be transmitted.

In the following, it is assumed that all data transmitted before the poll timer expires, are successfully received before handover, but that the status message has not reached the sender. This is not expected to occur often, but when it does occur, then the delay may be reduced if RLC PDU context transfer is used. Furthermore, it is assumed that there are always enough data available to fully utilise the radio link, and that the RLC transmission window is large enough not to get stalled. If only RLC SDU context transfer is applied, then the RLC state is reset due to handover and all forwarded RLC SDUs will be transmitted again to the UT, even though they were successfully received before handover. The time required to retransmit all outstanding data on the new path may be longer or shorter than the time to transmit the data over the old path, depending on the delay on the new path. In case of a BS_i-BS_jRN handover, for example, the delay would be longer on the new path, and in case of a BS_iRN-BS_j it would be shorter.

If the delay is the same as before, then it will take $T + d + d$ to retransmit the data and to exchange poll request and a status message. Thus, it takes $T + d + d$ before the RLC sender receives the status message about a successful transmission and can start to transmit new data. The delay may be reduced and unnecessary retransmission may be avoided, if RLC PDU context transfer is used and a status message is transmitted immediately after handover is completed. If the status message reaches the RLC sender, no retransmission is needed, since all outstanding data were successfully received already before handover. With RLC PDU context transfer, the RLC sender will receive

the status message about successful transmission after *d* instead, which is a shorter delay than with RLC SDU context transfer.

Figure 5-13 and Figure 5-14 show the outstanding data before handover as a function of the poll timer for different values of the bandwidth and the delay. Larger delays are handled better with more bandwidth. A delay of 5 ms is typical for a direct communication between a UT and a BS. A delay of 20 ms can be expected when there are RNs on the communication path. With a poll timer set to 100 ms, *b* = 10 Mbps, and *d* = 20 ms, the outstanding data is 700 kbits.

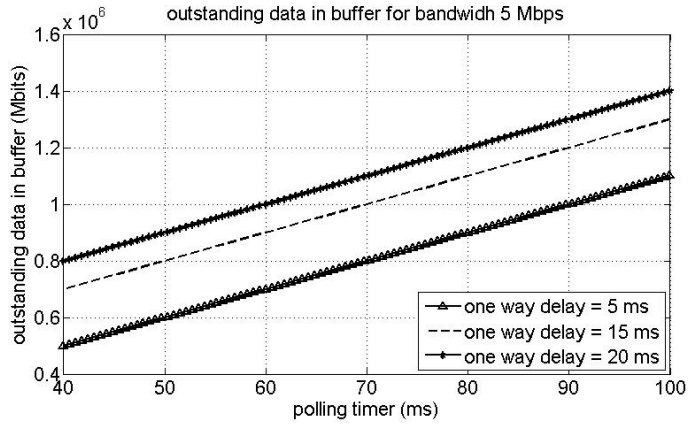


Figure 5-13 Outstanding data before handover for b= 5 Mbps and different delays.

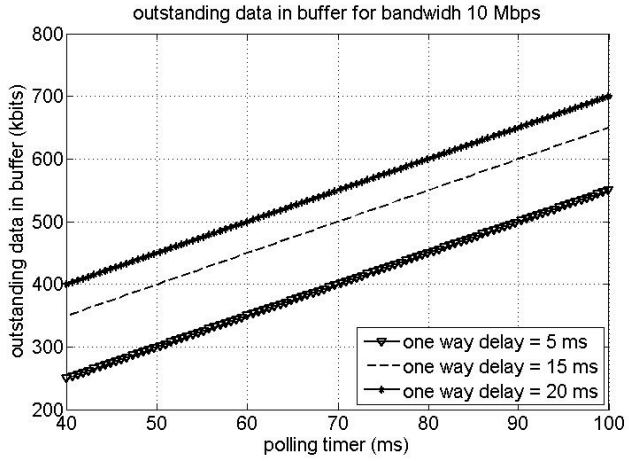


Figure 5-14 Outstanding data before handover for b= 10 Mbps and different delays.

Figure 5-15 shows the time before new data can be transmitted after handover. With only RLC SDU context transfer, transfer of new data could start only after 140 ms, after the forwarded but already received data are retransmitted to the UT (provided that the new path has the same characteristics). With RLC PDU context transfer, and assuming that a status message is transmitted immediately after handover, transfer of new data can instead starts already after 20 ms.

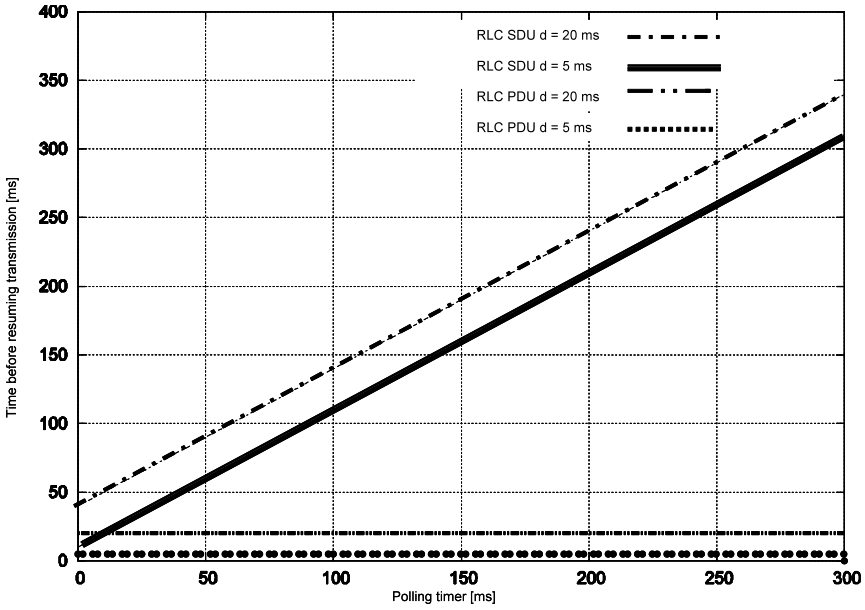


Figure 5-15 Time before new data can be transmitted after handover

Therefore, it is beneficial to employ RLC PDU context transfer for large amounts of data to be transferred. If only RLC SDU context transfer is applied, then the extra delay caused by the unnecessary retransmission of RLC SDUs may have a negative impact on the performance of higher layer protocols. New data is delayed, if the buffered RLC SDUs are retransmitted first. Depending on the relation between the delay and the TCP retransmission timer, this may lead to TCP retransmission timeout and that data will be unnecessarily retransmitted also by TCP.

For low amounts of buffered data RLC SDU context transfer would be sufficient.

5.3.3.3.2 In Sequence Delivery and Duplicate Detection

An important function for the IPCL is to provide in-sequence delivery of upper layer PDUs. In the DL, data may arrive out-of-order to the target BS, if the data transmitted directly from the GW to the target BS arrives before the data that are forwarded from the source to the target BS during handover. In the UL, the GW may receive data out-of-order, if there are gaps in the data transmitted from the source BS before handover and retransmissions of the missing data arrive from the target BS after handover [24].

Solutions to reorder out-of-order data are considered in newer versions of the LTE specifications [20].

In the context of the proposed RRM framework, performance degradation due to in order delivery would not be very severe because IPCL terminates in the GW and UT (see Figure 5-3). Out-of-order data, however, should be avoided whenever possible, because out-of-order data increases delay on higher protocol layers.

In [24] a separate service class for forwarded data is proposed. To reduce delay of forwarded data, the scheduler in the target BS is proposed to give priority to forwarded data. In [26] forwarded data is proposed to be prioritized over data from the GW in the target BS for transmission to the UT (which will only work if the buffer of forwarded data is not emptied before all forwarded data have been received). Here, it is proposed that forwarded data are prioritized, in order to avoid out-of-order data.

In the UL, the IPCL in the GW should remove duplicates before data are delivered to higher layers.

In the DL, there are two alternatives:

1. RLC in the target BS uses the IPCL sequence number to detect duplicated IPCL PDUs.
2. The IPCL in the UT detects and removes duplicated IPCL PDUs.

On one side, if alternative 1 is used and the RLC in the BS performs duplicate detection, then RLC has to look into the IPCL sequence numbers, which violates the protocol layering. On the other hand, if alternative 2 is used and IPCL in the UT performs duplicate detection, then the layering is preserved, but duplicates are transmitted all the way over the air to the UT. If there is enough capacity between the BS and the UT, then the UT could detect and remove duplicates and violation of protocol layering could be avoided. Duplicates are not expected to occur often. Therefore, it is recommended that IPCL in the UT performs duplicate detection. In environments, in which duplicates occur frequently, RLC in the BS could perform duplicate detection as a value added function.

5.4 Handover Priority Setting

According to the service profile the user registered, direction of handover can be controlled by the policy. Let user *A* and *B* be both subscribers. *A* subscribes to lower class; on the contrary *B* is rather a premium user. In that case, ranking of the handover candidate cell list can be differently ordered, e.g. user *A* has rank: LA cell *x* → LA cell

y → WA cell *z*; however user B is allowed to set handover list with ranking WA cell *z* → LA cell *x* –LA cell *y*. Context of A has also higher priorities than B to be transferred between neighbouring cells.

The handover priority setting helps the QoS guarantee for super class users. In this context a flow admission control function is proposed. This is a function, which grants or rejects requests based on the network resource availability and their priority. In broad terms, flow admission control limits the access to some resource such that the load on that resource remains limited. The actions of a flow admission control are shown in Figure 5-16.

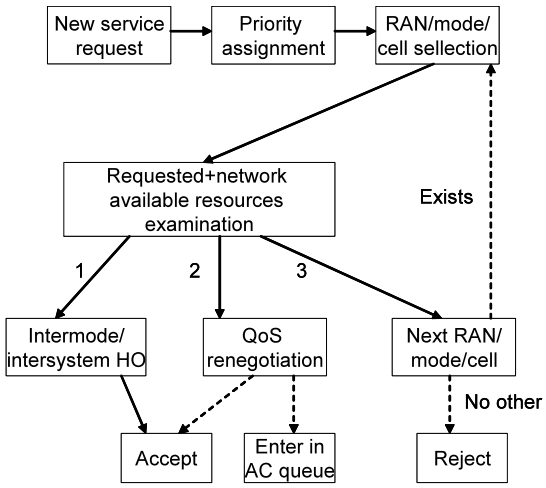


Figure 5-16 Flow admission control function

Further, a handover priority policy can be used in the case when there are several downlink connections (corresponding to the connections from BS to many UTs or to the connections of different services). In this case, the downlink TCP connection, which is about to perform handover process is given the highest priority. This connection is given highest priority by different known techniques depending on the systems, such as by allocating higher bandwidth, higher downlink bursts, or higher number of downlink subcarriers in OFDM(A) systems [5]. The technique in a way 'labels' the link layer segments to indicate their priority. This scheme can also be implemented to the uplink connection. The labels are based on the BS ID.

The required message exchange is shown in Figure 5-17.

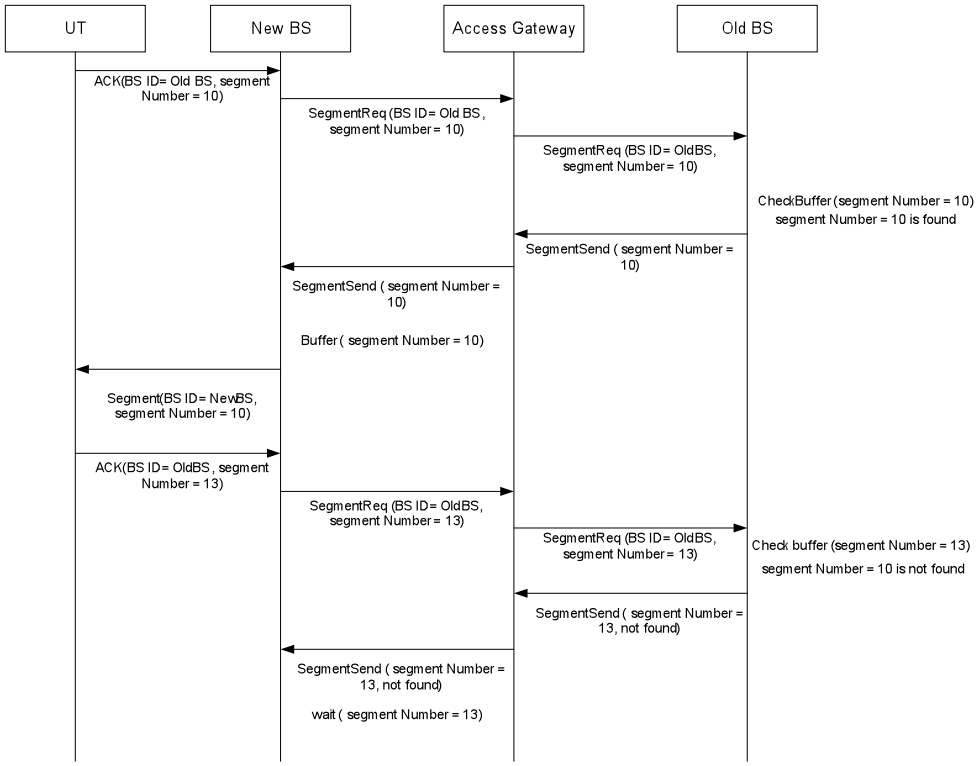


Figure 5-17 Message exchange for handover priority policy.

The purpose of this approach is to empty the buffers at the BS as soon as possible. Therefore, when handover happens, there will be no / small number of downlink TCP packets left in the BS buffer, hence reducing the number of TCP packet losses due to handover. This approach is suitable for multi-hop communication systems because it makes the sender aware that some of the occurred delays are not caused by congestions but by several hops on the transmission path.

This scheme is an event-driven based scheme, which means that the prioritization can be initiated (e.g. based on the signal strength value from the BS received by the corresponding UT).

5.5 Conclusions

Next generation systems offer new possibilities for advanced radio resource management. The flat architecture proposed for the RAN for IMT-Advanced candidate systems has been exploited to advance further the proposed concept of intra-system interworking between RRM components and the location of these functionalities. In particular, intra-system interworking can benefit from a combined centralized and distributed approach. The properties of a hybrid approach to intra-system RRM were used to propose strategies for the support of mobility management interactions (e.g., RAT/BS association) and as a means to improve the handover efficiency, for the

support of flow establishment and release as a means to improve the efficiency of user context transfer during intra-system handover. It was proposed that user context transfer is based on a mandatory and an additional functionalities and it was shown that it is useful to employ the additional functionality for selected cases which involve larger delays on the one-hop link. Future work envisions the enhancement of the proposed strategy to activate the additional functionality to flows based on their QoS characteristics. Finally, it was shown that for multi-hop communications, a labeling of the link layer transmissions by the proposed handover priority policy can prove beneficial to TCP performance for the case when several simultaneous downlinks exist. This approach is in line with the proposed physical layer characteristics (e.g., OFDMA) for IMT-A candidate systems.

References:

[1] E. Mino, A., Mihovska, et al., "D 4.8.1 WINNER II Intramode and Intermode Cooperation Schemes Definition," Deliverable D4.8.1, IST Project WINNER II, June 2006.

[2] A. Mihovska, et al., "Policy-Based Mobility Management for Next generation Systems," Proc. of IST Mobile Summit 2007, Budapest, Hungary, July 2007.

[3] A.-G. Acx, A. Mihovska, et al., "D1.3 Final Usage Scenarios," Deliverable 1.3, IST 2003-507581 Project WINNER, at www.ist-winner.org.

[4] Long Term Evolution, http://www.3gpp.org/Highlights/LTE/LTE.htm

[5] RECOMMENDATION ITU-R M.1645, "Framework and Overall Objectives of the Future Development of IMT 2000 and Systems Beyond IMT 2000," at www.itu.int.

[6] IST Project AROMA: Advanced Resource Management Solutions for Future All IP Heterogeneous Mobile Radio Environments, http://www.aroma-ist.upc.edu/.

[7] J., Postel, "Transmission Control Protocol," RFC793, http://www.ietf.org, Sept. 1981.

[8] S. Dixit and R. Prasad, *Wireless IP and the Mobile Internet,* Artech House Publishers, Boston, Ma., 2003.

[9] P., J., Ameigeiras, J., Wigard, and P., Mogensen, "Impact of TCP Flow Control on the Radio Resource Management of WCDMA Networks," in proc. IEEE 55th Vehicular Technology Conference VTC, May 2002, Vol. 2, pp. 977-981.

[10] M., Lott, " ARQ for Multi-Hop Networks," Proc. Of IEEE…., 2005.

[11] Seung-Gu Na and Jong-Suk Ahn, "TCP-like Flow Control Algorithm for Real-Time Applications," IEEE International Conference on Networks, ICON, September 2000, pp. 99-104.

[12] H., Balakrishnan, et al., "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links," IEEE/ACM Transactions on Networking, December 1997, Vol. 5, Issue 6, pp. 756-769.

[13] D., W., Browning, "Flow Control in High-Speed Communication Networks," IEEE Transactions on Communications, July 1994, Vol. 42, Issue 7, pp. 2480-2489.

[14] A., Klockar, A., Mihovska, et al., "Network-Controlled Mobility Management with Policy Enforcement towards IMT-A," *Proc. of ICCCAS 2008*, held on May 25-28, 2008, Xiamen, China.

[15] E., Mino, A., Mihovska, et al., , "D4.8.3 Integration of Cooperation in WINNER II System Concept," Deliverable 4.8.3 IST Project WINNER II, November 2007, www.ist-winner.org.

[16] Prasad, R., Mohr, W., and W. Konhäuser, Third Generation Mobile Communication Systems, Artech House 2000.

[17] F., Lefevre, G., Vivier, "Optimizing UMTS Link Layer Parameters for a TCP Connection," in Proc. of IEEE VTS 53rd Vehicular Technology Conference, VTC, May 2001, Vol. 4, pp. 2318 – 2322.

[18] M., Malkowski, and S., Heier, "Interaction between UMTS MAC Scheduling and TCP Flow Control Mechanisms," International Conference on Communication, ICCT, April 2003, Vol. 2, pp. 1373 – 1376.

[19] 3GPP TR 25.813 (2006-06), "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Radio Interface Protocol Aspects," at www.3gpp.org.

[20] 3GPP TS 36.322 V8.1.0 (2008-03), "Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Link Control (RLC) protocol specification (Release 8)," March 2008 at www.3gpp.org.

[21] S., Jain, and E., Modiano, "Buffer Management Schemes for Enhanced TCP Performance over Satellite Links," at www.mit.edu/~modiano/papers/C90.pdf

[22] R. Guerin, et al., " Scalable QoS Provision through Buffer Management," in Proc. of SIGCOMM 1998, at www.sigcomm.org/sigcomm98/tp/paper03.pdf.

[23] R. Zurawski, Editor, The Industrial Communication Technology Handbook, Chapter 1, CRC Press 2005, Taylor&Francis, Boca Raton, Fla.

[24] L., Bajzik, et al., "Impact of Intra-LTE Handover with Forwarding on the User Connections", Proc. of the 16th IST Mobile Summit 2007, Budapest, Hungary, July 2007.

[25] J., Sachs, et al., "Evaluation of Handover Performance for TCP Traffic Based on Generic Link Layer Context Transfer," Proc of. IEEE 17th International Symposium on PIMRC 2006, September 2006.

[26] A., Racz, et al, "Handover Performance in 3GPP Long Term Evolution (LTE) Systems", Proc. of the 16th IST Mobile Summit 2007, Budapest, Hungary, July 2007.

[27] J., Alcaraz, et al., "Optimizing TCP and RLC Interaction in the UMTS Radio Access Network," *In IEEE Journal on Networking 20(2):* 56-64, March/April 2006.

[28] Y.-Ch., Chen, et al, "Simulation Analysis of RLC for Packet Data Services in UMTS Systems," in Proc. of IEEE PIMRC 2003.

Chapter 6

Multi-Stage Admission Control

This Chapter proposes and evaluates a novel multi-stage distributed admission control algorithm based on the proposed in Chapter 2-Chapter 5 intra-system cooperative RRM framework. In particular, the interworking between RRM functionalities for congestion, admission and load control located at the RN, BS, GW, and the RRMServer are exploited. The proposed multi-stage admission control mechanism is also suitable for a multi-hop communication system.

The proposed algorithm uses the concept of hybrid RRM to provide for load balancing and faster response time to admission requests of users in a next generation system. In a scenario, where a high throughput demanding session/call is subject to be admitted, it would not be sufficient only to check the load or capacity within the RAN, since the backhauling or the core network might be the bottleneck. Backhaul lack of resources has been considered in [1], [2] for the cell selection process. There, it was claimed that resource limitations in the transport network may result in blocking of new sessions and/or service degradation in terms of delay and packet error rate during an overload. The hybrid RRM approach proposed earlier allows that the multi-stage admission control takes into account the available in the backbone resources together with the available in the RAN resources for a decision based on the measured load.

The proposed admission control mechanism is load-dependent and it uses decision polling based on the load in entities located at different levels of the RAN architecture hierarchy. Thus it is based on information sharing at different levels of the RAN architecture. Load sharing is an important technique that improves distributed system performance by letting a group of entities in a system share their performance [3]. By load sharing, a better utilization of resources at all the entities of the RAN, and consequently, a high system throughput or short response time of user requests can be achieved. Most of the load-sharing algorithms are designed for tightly coupled distributed systems [4].

Load sharing has been exploited in single layers [5] and between mobile systems for reducing the number of unnecessary handovers for already connected users. Traditionally, load sharing algorithms are based on the token-passing paradigm.

Here, it is proposed that the load information is shared by passing a token between the RNs, BS and the GW (see Figure 2-1), thereby, considering the load status at cell and system level. A token is assigned to the entity with the highest load, this entity becomes the token holder who would be the one to respond upon an admission request. Each admission request to a token holder will issue a flag that reflects the load level in this entity, which in turn would activate a sequence of admission control actions related to the admission of a user to the network, i.e., the ranking of the intermediate decisions is dynamic.

The proposed mechanism aims at distributing as much as possible the radio load among the serving entities so that users can be more uniformly distributed in the network and served at an acceptable QoS.

This Chapter is organised as follows. Section 6.1 describes the scenarios for the multi-stage admission control. Section 6.2 gives the mathematical framework for the proposed mechanism. The proposed multi-stage admission control is implemented and evaluated in terms of faster response to user admission requests and gain from load balancing. The assessment is given in Section 6.4. Section 6.5 concludes the Chapter.

6.1 Scenarios for Multi-Stage Admission Control

The scenario for multi-stage admission control is shown in Figure 6-1.

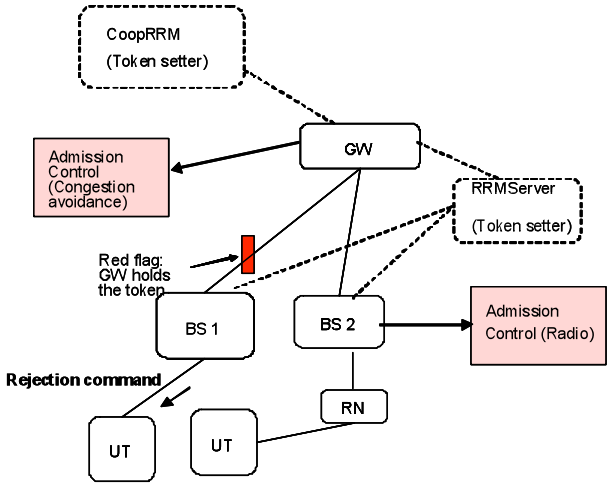


Figure 6-1 Token setting entities for multi-stage admission control.

It assumes two types of token setting entities:

1. Token setters for situations of low loads based on the distributed RRM framework (indicated by firm lines). In this case the decision framework includes the GW, BS, and RN.
2. Token setters for situations of medium to high load based on the centralized RRM framework (indicated by dotted lines). In this case the decisions are taken by the CoopRRM and RRM Server.

The proposed multi-stage admission control framework has the advantage of preserving the individual behaviour of the entities but exploiting the benefits for information sharing for ensuring load balancing through interworking between these entities.

6.1.1 Token Setting for Sequential Flag for Single-Hop

When token setting is applied in a single-hop scenario, the admission control function is distributed among the BS and the GW. The BS takes care of the radio part during admission control, while the GW takes care of congestion avoidance within the core network or other sub-networks.

If assumed that the decision made by the BS is **D**1, and the decision performed by the GW is **D**2, then a user *N* will be rejected when $\mathbf{D}_i = 0$ and accepted when $\mathbf{D}_i = 1$, with $i = 1$ or 2. The final decision for the incoming call will be a Boolean operation:

$$D_{\mathbf{final}} = \mathrm{D}\mathbf{1}\ \mathbf{AND}\ \mathrm{D}\mathbf{2} \qquad \textbf{(6-1)}$$

Admission control is not immediately performed if a *'green flag'* is received; instead this case will require further checking of the most instantaneous situation. In the case of a *'red flag',* a rejection command is immediately issued without checking the available capacity or bearer. A *'yellow flag'* will be sent when a shared resource can be repartitioned. This mechanism is shown in Figure 6-1 for a sent *'red flag'*. In this case, the GW holds the token, and after identifying that no resource is available in the backbone it sends a *'red flag'* to the BS, to which an admission request is made, and it in its turn issues the rejection command.

Along with the rejection command, the GW may provide other candidate BSs information of the BSs pool interconnecting to the GW. In this way a three-fold benefit can be achieved: load balancing among the BSs, a higher user satisfaction, since the rejection will be a *'soft'* command, and a relaxation of the load on the I_{BB} interface.

The distributed AC functions can be ranked according to different criteria. The token will determine the sequence of the AC depending on the criteria upon which the tokens have been set. One criterion is the *system load*. For the scenario in Figure 6-1 the current load of the RAN is low, and the BS has relatively high capacity. At the same time, the GW identifies a higher probability in congestion in the backbone or a limitation from other sub-networks, which the expected traffic has to go through. In that case, the GW holds the token to perform the congestion prediction first before the radio admission control is performed. The token assigned to the GW will be ranked higher.

The token assignment can be also *service dependent.* For example, if an incoming high rate data service (e.g., high FTP) needs different token assignment, such as voice like service, it would require an early check at the GW (i.e., high data rate services can cause congestion on backhaul), while low data rate real-time service would require an early check at the BS.

If an incoming session is expected to add too much load in terms of data rate to the backbone network or to other sub-networks, the GW will send the *'red-flag'* to the corresponding BS where the UT is connecting to. The BS then immediately rejects the session.

If the GW identifies that the capacity in the backbone is sufficient, it would issue a *'green flag'*. After receiving the *'green-flag'*, the BS would check if the available resources are sufficient for the incoming call, and admit the user if this is the case. An alternative sequence of events will include a cell reselection admission control algorithm or resource repartitioning.

6.1.2 Token Setting for Sequential Flag for Multi-Hop Scenario

In the multi-hop scenario shown in Figure 6-2 the admission control mechanism takes into account the intermediate relaying capacity between the RNs and between the RN and the BS. The procedures are pretty much identical to the ones explained for the single-hop scenario. For the multi-hop scenario it is necessary to check the available capacities of all segments of all possible routing paths between the UT and the BS before an admission command is generated. The segments are defined according to the direct connections among the RNs and the BS. The token holder is given considering a downlink (DL)/uplink (UL) transmission and the RN closest to the to-be-admitted UT. This is shown in Figure 6-3.

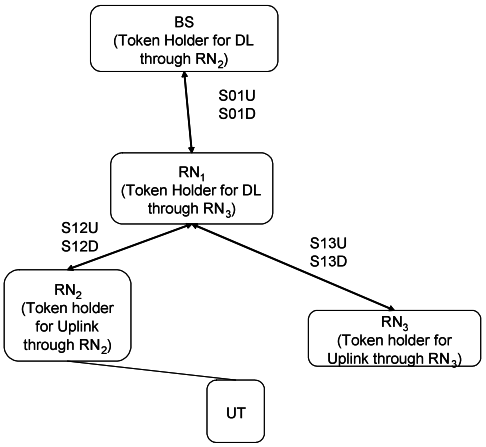


Figure 6-2 Segments and token holder in a multi-hop scenario.

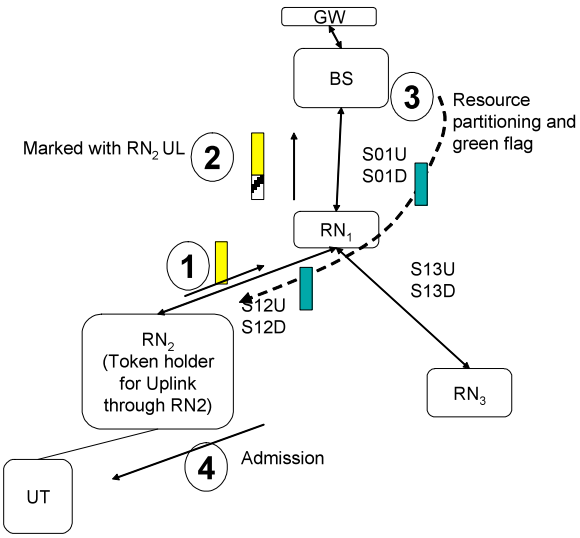


Figure 6-3 Yellow-' (Soft)-flag and piggy-packed 'green flag' along with resource partitioning.

For an FDD system [6], the capacities of the UL and DL will be different, which makes the token holder assignment also different. For example, in the poorest link in the UL from the RN2 point of view is **S12U**, therefore the token is assigned to RN2; however, in the DL, the poorest link is **S01D** respective to RN2, therefore, the token is assigned to BS.

When RN2 identifies a lack of resource for the incoming calls (bottleneck identified or Step 1), it sends immediately a *'yellow-flag'* (soft-flag) with its marker/header to the central resource control unit in the cell (typically the BS) as Step 2 in Figure 6-3. The BS checks with the GW about the core network resource and

repartitions the resources and at the same time confirms to the RN2 by sending the *'green flag'*. RN2 then admits the UT when it confirms that the resources are sufficient.

In this case, the resources among the involved entities are shared. The gain achieved by load sharing is analysed later in this Chapter. For each RN, the optimal routing paths for any potential incoming sessions can be restored. Throughout the path, the bottleneck will be identified according to the QoS expected from the incoming calls (service context). The *'red'* or *'green flag'* goes always to the next decision maker. However, the *'yellow-flag'* (soft-flag) is among the coupled entities that may perform a resource repartitioning in order to allow the incoming calls.

6.2 Gain Analysis for Load Sharing

For the multi-stage admission control, load sharing is analysed between cooperative intra-system entities (e.g., GWs, BSs and RNs). First an expression for the load is derived that is common to all entities.

6.2.1 Derivation of Load Definition

To obtain an expression for the load that would be common to all entities involved in the multi-stage admission control, it is proposed that the cell/system load can be modelled as an exponential distribution with a parameter called weight of noise rise, modelling the sensitivity of mutual interference or that Equation 6-2 holds:

$$L = e^{an} \qquad \textbf{(6-2);}$$

where n is the number of traffic units added to the network; and a is the weight of the noise rise. Different deployment modes have different noise rise curves due to the properties given by the physical modes [6].

Noise sources (e.g., UTs) may be characterised by the maximum amount of noise power or power spectral density that can be passed to a load [7]. As an example, the users' transmission power increase due to a new user in a CDMA cellular system depends directly on the current uplink noise rise [8]. This is also the reason why the load, L, can be related to the noise rise. A dependency is also given by the *pole equation* [9], [10].

Suppose the noise power density is σ_N^2 with the bandwidth W, then the noise power is as in Equation 6-3:

$$P_N = \sigma_N^2 \cdot W \quad \textbf{(6-3)}$$

A parameter called *noise rise* can be named, such as:

$$\zeta = \frac{I}{P_N}$$

where the *interference I* is defined by:

$$I = P_N + \alpha \cdot \sum_j P_j \quad \textbf{(6-4)}$$

where $\alpha << 1$ is a coefficient that calculates the actual interference and *j* is the index for the users in the service area each transmitting with power P_j.

If S_j is adopted as the SNR for user *j*, then it can be calculated that

$$S_j = \frac{P_j}{I - P_j} \Rightarrow I \cdot S_j - P_j \cdot S_j = P_j \Rightarrow P_j = \frac{1}{1 + \frac{1}{S_j}} \cdot I \quad \textbf{(6-5);}$$

The load factor for user *j* is given by:

$$L_j = \frac{1}{1 + \frac{1}{S_j}} \quad \textbf{(6-6)}$$

By combining Equation 6-6 and Equation 6-7 together, the noise rise definition in Equation 6-4 can be redefined to include the load value as:

$$\zeta = \frac{1}{I - \alpha \cdot \sum_j P_j} \Rightarrow \zeta = \frac{1}{1 - \alpha \cdot \sum_j L_j} \quad \textbf{(6-7).}$$

If $\eta = \alpha \cdot \sum_j L_j$ exists, then the interference increment can be obtained as:

$$\zeta = \frac{I}{P_N} = \frac{1}{1-\eta} \Rightarrow \frac{dI}{d\eta} = \frac{P_N}{(1-\eta)^2} \Rightarrow \Delta I \approx \alpha \cdot \Delta L \cdot \frac{P_N}{(1-\eta)^2} \Rightarrow \Delta I \approx \alpha \cdot \Delta L \cdot \frac{I_O}{1-\eta} \quad \textbf{(6-8)}$$

The *weight of noise rise* can be defined as

$$\beta = \frac{\alpha}{1-\eta} \quad \textbf{(6-9),}$$

such that the system load can be modelled as simple exponential function. Therefore, in case the number of links increases by 1, Equation 6-9 can be rewritten as:

$$I(n) - I(n-1) \approx \beta \cdot I(n-1) \quad \textbf{(6-10)}$$

If it is assumed that the system load can be modelled by interefence as an exponential function or that:

$$I(n) = e^{\gamma n} \quad \textbf{(6-11);}$$

where γ the *weight of noise rise,* and n the number of current users/radio connections.

Using Taylor series, Equation 6-12 can be rewritten as:

$$I(n) = e^{\gamma(n-1)} + \gamma \cdot e^{\gamma(n-1)} \cdot [n-(n-1)] + \frac{\gamma^2}{2!} \cdot e^{\gamma(n-1)} [n-(n-1)]^2 + \ldots$$
$$\Leftrightarrow I(n) \approx I(n-1) + \gamma \cdot I(n-1) + o(I(n)) \quad \textbf{(6-12)}$$

It can be seen that system load can be modelled by exponential increase by introducing a weight of noise rise factor depending on the number of users/radio links. This means that the definition of Equation 6-1 holds.

6.2.2 Gain from Load Sharing

The gain analysis assumes that on the average the number of active calls remains constant and that each BS has the same capacity. The decision load is defined as the admission request intensity per time unit.

The model of the load (related to noise rise) for a single cell was given by Equation 6-2.

Suppose that the load contribution (additional load) can be distributed over two GWs/BSs/RNs, or $m = m_1 + m_2$, where m is the amount of incoming traffic units to an entity. For the individual BS, the load increase of the cell where all the traffic is added to is then given by:

$$L_i' = e^{a(n_i+m)} - 1 \quad \textbf{(6-13)}$$

with i the entity index. In the following description, superior ‘′’ denotes that the added traffic is not distributed, i.e. the incoming traffic with units m is only added to a single BS; superior ‘″’ is used to show that the load is balanced over two entities.

For the traffic distribution case, the load values in the entities are calculated as:

$$L_1'' = e^{a(n_1+m_1)} - 1 \quad \text{and} \quad L_2'' = e^{a(n_2+m_2)} - 1 \quad \textbf{(6-14).}$$

Let $n_1 = n_2$ and the traffic is added to a single entity. Without losing the generality, we assume that entity 1 receives all the added traffic, its load is therefore as in Equation 4-7:

$$\Delta' = L'_1 - L_1 = (e^{a(n_1+m)} - 1) - (e^{an_1} - 1) \quad \textbf{(6-15)}$$

For the case that traffic is added to both entities, the load increase is given by:

$$\Delta'' = \sum_i (L''_i - L_i) = \sum_i \left[\left(e^{a(n_i+m_i)} - 1 \right) - \left(e^{an_i} - 1 \right) \right] \quad \textbf{(6-16)}$$

with $i = 1,2$. With the assumed condition $n = n_1 = n_2$, $m = m_1 + m_2$ and $m_1 = m_2$, the ratio between the added loads for the load sharing scenario and the single cell scenario is as in:

$$R = 2 \cdot \frac{e^{a(n+m/2)} - e^{an}}{e^{a(n+m)} - e^{an}} \quad \textbf{(6-17)}$$

Finally, the gain for load balancing is obtained as by:

$$G = \frac{1}{2} \cdot \frac{e^{a(n+m)} - e^{an}}{e^{a(n+m/2)} - e^{an}} - 1 = (e^{a \cdot m/2} - 1)/2 \quad \textbf{(6-18)}$$

As $a > 0$, and $m/2 > 0$, so that $G > 0$.

Figure 6-4 to Figure 6-8 show the results based on the gain analysis described above for different values of *n* and *m*, where the contribution to the load in traffic units is expressed as the average arrival rate of users at the BS or at the GW pool of BSs.

For all values of *m* and *n*, there is a gain from load balancing as compared to the results for the case when a single BS takes all the incoming traffic. Increase of the value of *m* does not change the effect from the proposed algorithm. In case when the average arrival rate is very low (i.e., *n* = 0.5), the load is not very high even without the use of load balancing, which is visible from the graph in Figure 6-4.

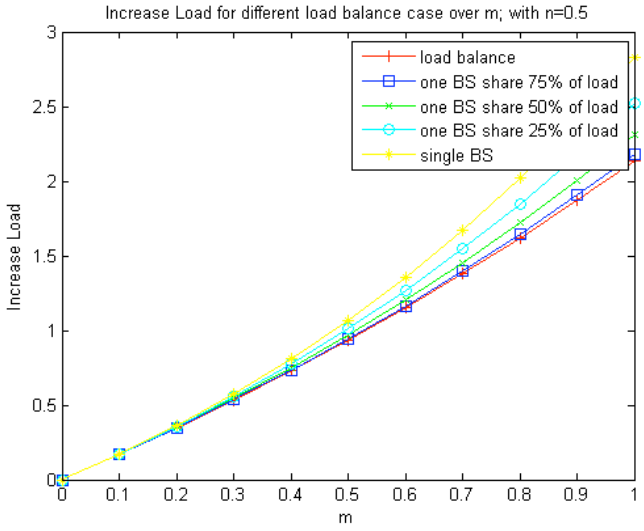


Figure 6-4 Effect of load balancing for m= 1, n= 0.5

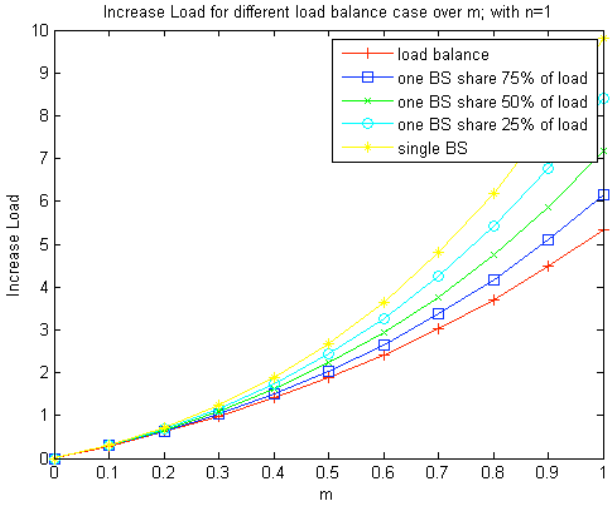


Figure 6-5 Effect of load balancing for m= 1, n = 1

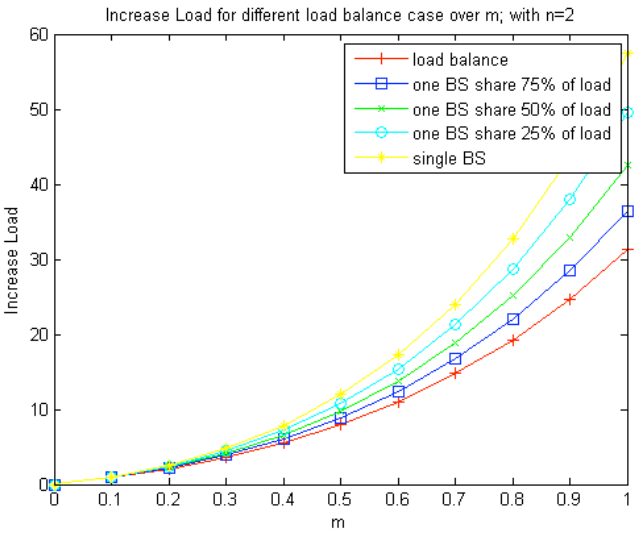


Figure 6-6 Effect of load balancing for m=1, n = 2

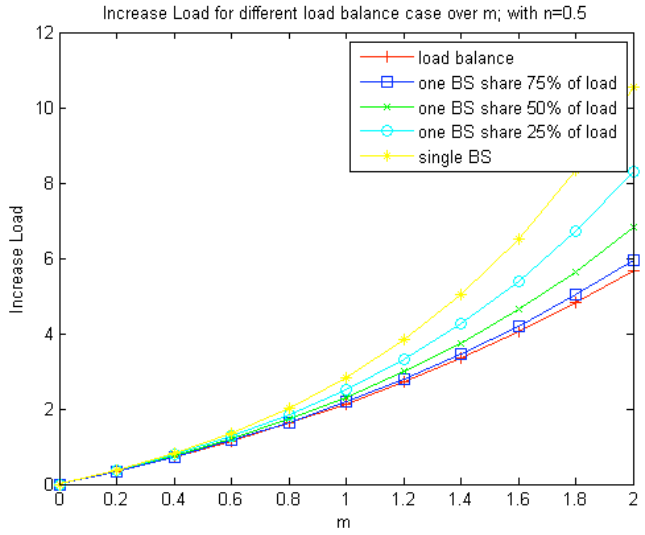


Figure 6-7 Effect of load balancing for m = 2, n = 0.5

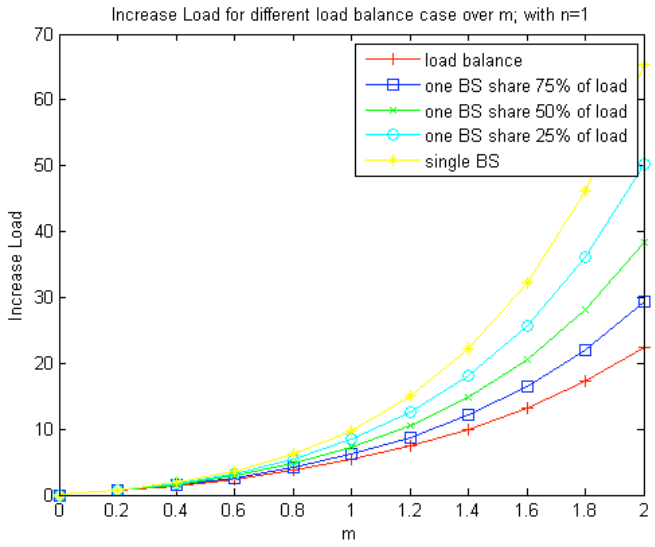


Figure 6-8 Effect of load balancing for m = 2 and n = 1

The effect from load balancing is more obvious in a case when the average arrival rate of users to a serving area is higher. It can be seen that the gain from load balancing is higher for larger values of *n* which leads to an overall load increase for a single BS of at least 20% when no balancing is applied.

6.2.3 Gain from Interworking between BSs

Further gain can be obtained from interworking between entities (e.g., BS and RNs), Such gain is referred to as multiplexing gain and it is achievable through the proposed hybrid RRM framework for cooperation which uses the advantages of a centralised and

distributed RRM, in particular related to scheduling of users after a load sharing decision.

Assuming that all users have the same QoS demands, irrespective to the user queue, by assigning the resource (i.e. ensuring radio access) to the user with best channel capacity, multiuser diversity gain can be obtained [11]. Such gain is obtainable by employing the mechanism of joint radio resource management for the scheduling as proposed in [12]-[16].

The more resources are available for a system, the higher the multiuser diversity gain that can be obtained.

Theoretically, it can be assumed that for all the users to be scheduled a *Signal to Interference plus Noise Ratio (SINR) density function* can be defined as $f_S(s)$. The complimentary *Cumulative Density Function (CDF)* for a SINR threshold S_T then can be defined as:

$$\overline{F}_{S_T}(s) = \int_{S_T}^{\infty} f_S(s)ds \qquad \textbf{(6-19)}$$

If assumed that a total of N users is controlled by the scheduler, then the probability that at least one user will have a *CDF* higher than the threshold value is:

$$\sum_{i=1}^{N}\binom{N}{i}\overline{F}_{S_T}^{\,i}(s)\left[1-\overline{F}_{S_T}(s)\right]^{N-i} \qquad \textbf{(6-20)}$$

If the SINR is assumed Rayleigh distributed without co-channel interference and shadowing effect for all UTs, then the multiuser diversity gain can be derived from the probability distribution plot. The obtainable gain plotted for four UTs under the above assumptions is shown in Figure 6-9.

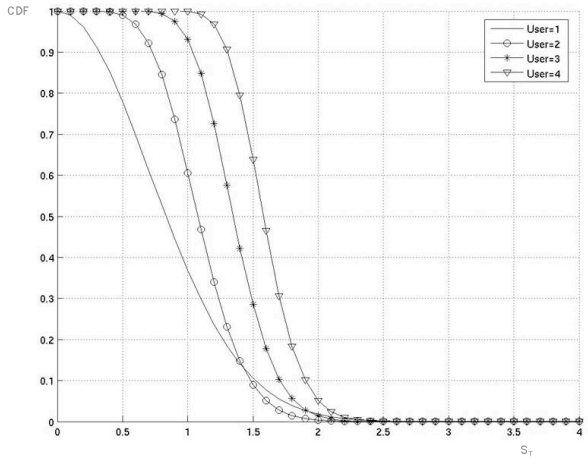


Figure 6-9 Multiuser diversity gain for four UTs.

In the following it is exploited that when a BS is operating on its own, it can be viewed as a single server [16]. Then it can assign the radio resources in terms of a small period of time (i.e. time transmission interval, TTI) to a UT when it needs to be served.

For a system with a single server, the average response time for a UT that requires service time *x* can be derived as:

$$T(x) = x/1 - \rho \qquad \textbf{(6-21);}$$

where $\rho = \lambda / E\left(\frac{C}{D}\right)$, C indicates the capacity offered by the system, D is the data amount to be transmitted by the service, λ is the arrival rate for the service controlled by the BS [17].

If *N* BSs are interworking at a scheduling level with each other and if the packets can be scheduled at each TTI to the UTs controlled by all of them, then, the average response time for a joint scheduling $T^{(J)}$ is given by:

$$T^{(J)} = 1 / \sum_i \left(\frac{C_i}{D} - \quad _i \right) \qquad \text{with} \quad i \in \{1, 2, ..., N\} \qquad \textbf{(6-22);}$$

When the *N* BSs are not able to interwork at the scheduling level, the response time is the average of the response time of each BS, or the response time for the case of non-joint scheduling $T^{(N)}$ is defined as in:

$$T^{(N)} = \frac{1}{N} \sum_i \frac{1}{\lambda C_i / D - \quad _i} \qquad \textbf{(6-23).}$$

Due to the nature of the response time, the denominators in both Equation 6-23 and Equation 6-24 must be positive, therefore a condition is defined that $T^{(J)} < T^{(N)}$.

The gain thus obtained is achieved by multiplexing of the resources from the interworking *N* BSs, therefore, this gain is termed as *multiplexing gain*. It allows toallocate the managed radio resources to the involved traffic in order to minimise the overall system load.

In terms of load balancing, this gain can provide benefits to the distribution of the SINR in interference constrained environments. These benefits are shown in Figure 6-10 for the SINR distribution of an OFDMA-TDMA based system where the load indicates what percentage of the resources in the coverage area are utilised on average. In one case the system uses about 20% of the available resources, in the other case, 80% of the available resources are in use.

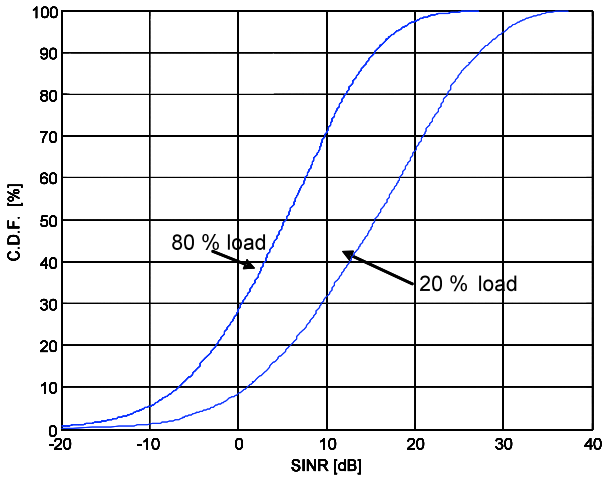


Figure 6-10 Effect on SINR distribution from load sharing.

From Figure 6-10 it can be seen that higher loads (when load balancing is not employed) lead to wider distribution of the SINR. With gain from load balancing through multiplexing gain, resources are shared in a balanced way because the response time for serving the users is reduced. At the same time less users are assigned to each entity (e.g., BS)

The performance of the system in general can be further improved by use of techniques, such as multiple input multiple output (MIMO) [18], adaptive modulation and coding, interference mitigation techniques [6] and so forth.

6.3 Implementation for the Token Setting

It is proposed to organise the RRM entities (see Figure 6-1) in a logical ring. The proposed implementation is restricted to the single-hop scenario (i.e., RNs are not included). The entities are interrogated (load is polled) every T_u seconds. As this scenario corresponds to resource partitioning within the RAN (intra-RAN resource partitioning), the time for allocating the resources will be in the order of seconds. Therefore, the polling time can be assumed in this order. Addition, removal and handoff of users will reorganise the load values for each entity and this will affect the token rotation (and change the values of the token rotation table.) The flowchart for the token passing is shown in Figure 6-11.

The default token can be set during the network planning and the maintenance phase. When the load and capacity change in the system, the default token can be

reallocated to the most severe entity following the principles/mechanisms described in the previous sections.

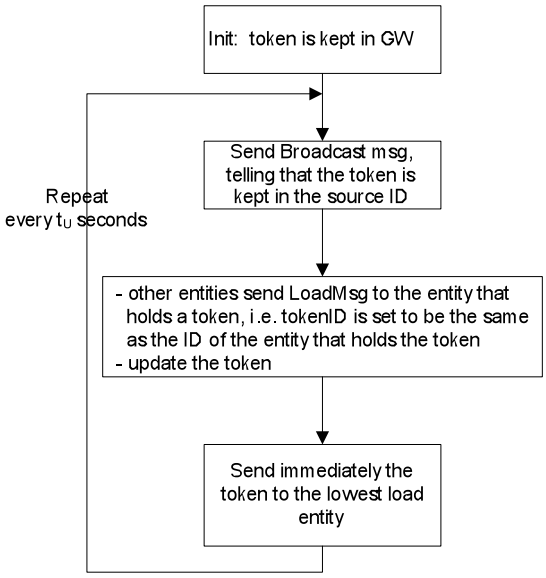


Figure 6-11 Flow chart for the token passing mechanism.

For a scenario of a GW pooling three BS together (see Chapter 2), the token passing can be implemented as shown in Figure 6-12.

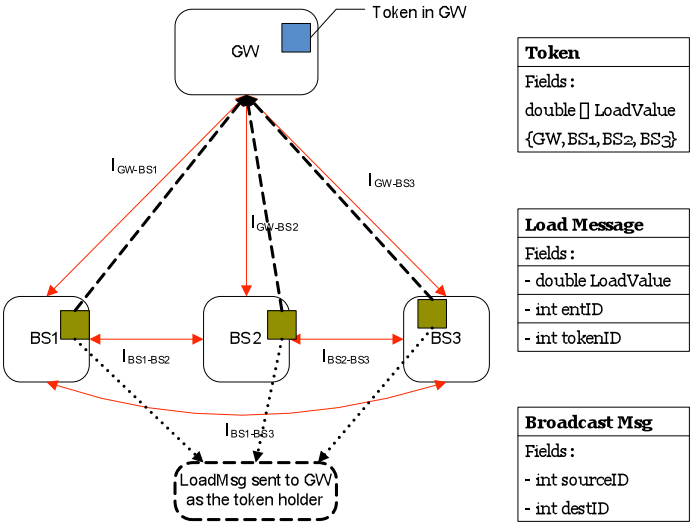


Figure 6-12 Entities and interfaces that involved in the token passing; and message fields of *LoadMsg* and *TokenMsg*.

The update of the load information is shown in Figure 6-13.

If the load in the BS is lower than that in the GW then the token is assigned to the BS. The applied rule can be expressed as follows:

$$Load_{BS} < Load_{GW} \rightarrow Token_{BS} \quad \textbf{(6-24)}$$

In this case, the BS will directly issue a token to GW whether it is a *green, yellow,* or *red* flag.

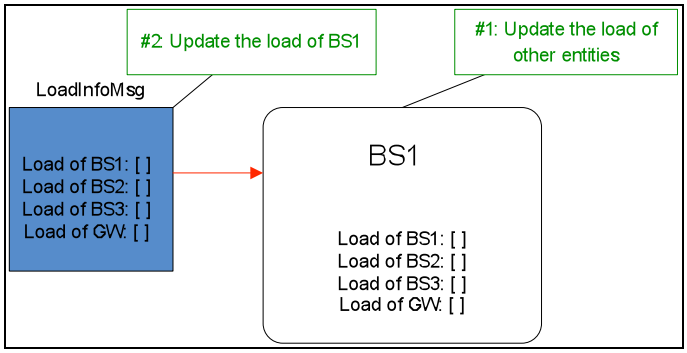


Figure 6-13 Implementation of the *LoadInfoMsg*.

If the load in the BS is higher than the load in the GW, then the token is assigned to the GW. The rule becomes as follows:

$$Load_{BS} > Load_{GW} \rightarrow Token_{GW} \qquad \textbf{(6-25)}$$

In this case, the BS will just compare with the value *Threshold_Load,* and produce a **D**1. Because the BS does not issue any flag or token, when it is forwarded to the GW, the GW already knows that a token or a flag, must be issued.

Figure 6-14 and Figure 6-15 show the flow charts of the admission control processes for the cases when the token holder is the GW and the BS, respectively.

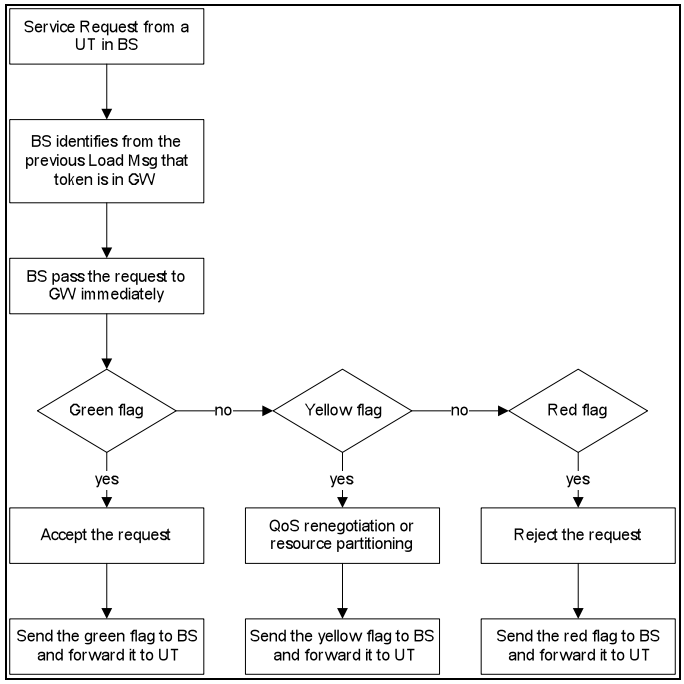


Figure 6-14 Admission control sequence when the token holder is the GW.

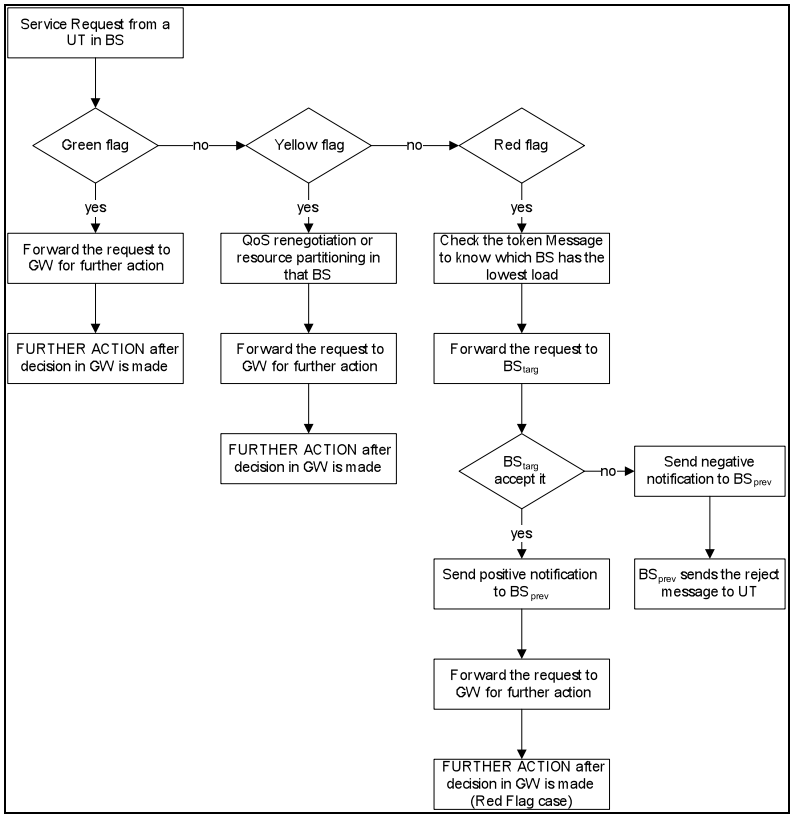


Figure 6-15 Admission control sequence when the token holder is the BS.

Figure 6-16 show an example of the follow up action after a decision is made in the case when the token holder has been the GW.

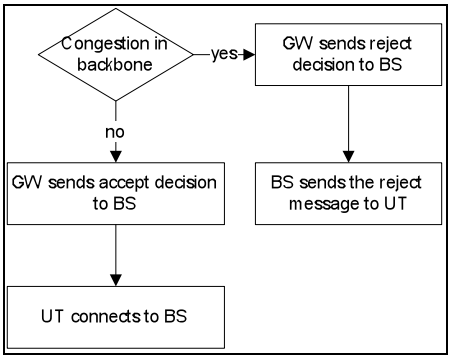


Figure 6-16 Follow up actions on a decision by the GW.

6.4 Conclusions

This Chapter proposed and analysed an algorithm that executes a load-dependent sequence of decision polling, i.e., the ranking of the intermediate decisions is dynamic. The algorithm considers that even if the decision is a successful admission in one part of the network (e.g., RAN) it might result in congestion in the core network. For a

scenario of a relay-enhanced network with multi-hop, where a wireless backbone provides connectivity and the far hops may not provide sufficient capacity for each segment of the end–to–end routing path, the proposed multi-stage admission control has benefits both for users by providing them with QoS and for network owners.

The proposed mechanism provides for a balanced load of the involved entities. Due to the balance, each entity has a lower load; therefore a potential decrease of the response time of the network entity can be obtained. (e.g., 10 requests per second in classic solution can be reduced to 5 requests per second to one entity). With use of the multi-stage admission control, hectic inter-GW-BS context transfer is avoided. For example, for new calls or sessions to be admitted, an acknowledgement (ACK) must be received through the I_{GB} and I_{WU} interfaces. This is an advantage for the case when the GW is the limiting factor, then, a traditional admission control performed only at the BS might result in a biased/wrong decision, which in turn will trigger user context transfer reallocation from the GW to the BS. There is also a potential reduction in air interface signaling.

In the proposed admission control architecture the interface between the BS (I_{BB}) is key to providing inter-BSs control and negotiation functions, like active mode mobility, interference management, spectrum functions, and load balance. It is for further work to investigate what information can be beneficial to signal over this interface to optimise the proposed multi-stage admission control.

If we view the radio system as a finite-state machine assuming a Poisson distribution of user arrivals and an exponential user service time, the trunking gain for can be obtained by a multi-dimensional Markov model. Such investigation can be beneficial to optimizing the proposed multi-stage admission control mechanism and extending the investigation to a scenario of inter-system cooperation. This mathematical framework is envisioned as part of the planned future work.

References:

[1] R., Ferrus, J., Olmos, and H., Galeana, "Evaluation of a Cell Selection Framework for Radio Access Networks Considering Backhaul Resource Limitations," in *Proc. of PIMRC'07,* September 2007, Athens, Greece.

[2] J., Olmos et al., "A Functional End-to-End QoS Architecture Enabling Radio and IP Transport Coordination," in *Proc. of WCNC'07,* March 2007.

[3] J., Cao, X., Wang, and S., K., Das, "A Framework of Using Cooperating Mobile Agents to Achieve Load Sharing in Distributed Web Server Groups," Elsevier Journal, October 2003, doi:10.1016/S0167-739X(03)00175-4.

[4] S., Krueger, et al., "Load Sharing Policies in Locally Distributed Systems," IEEE Computers, 1992, Vol. 25, Nr. 12, pp.33-44.

[5] A., Tölli, P., Hakalin, and H. Holma, "Performance Evaluation of Common Radio Resource Management," in *Proc. of ICC'02*, 2002.

[6] IST Project WINNER II Deliverable 4.7.1 "Interference Averaging Concepts," June 2007, at www.ist-winner.org.

[7] L., W., Couch II, *Digital and Analog Communication Systems,* Macmillan Publishing 1993.

[8] E., G., Lundin, et al., "Uplink Load and Link Budget with Stochastic Noise Rise Levels in CDMA Cellular Systems," Dept of Electrical Engineering, Linkoeping, Sweden.

[9] H., Holma and J., Laakso, "Uplink Admission Control and Soft capacity with MUD in CDMA," in Proc . of IEEE VTC 1999, Amsterdam, The Netherlnads, September 1999.

[10] R., Padovani, et al., "CDMA Digital Cellular: Field Test Results," in Proc. of IEEE VTC 2007, Stockholm, Sweden.

[11] E., Mino, A., Mihovska, et al., "Scalable and Hybrid Radio Resource Management for Future Wireless Networks," Proc. of IST Mobile Summit 2007, Budapest, Hungary, July 2007.
[12] J., Luo, and X., Huang, "Planning Future Heterogeneous Wireless Networks," *in Proc. of Symposium on Progress in Electromagnetics Research,* Hangzhou, China, August 2005
[13] J., Luo, et al., "Performance Investigations of JRRM in a Reconfigurable Environment," *in Proc. of SCOUT Workshop,* Paris, France, September 2003.
[14] J., Luo, et al., "Investigation on Radio Resource Scheduling in WLAN Coupled with 3G Cellular Networks," *IEEE Communications Magazine,* June 2003.
[15] IST 2003-507995 Project End-to-End Reconfigurability (E2R), Deliverable D5.3, "Algorithms and Performance, Including FSM& RRM/Network Planning," June 2005.
[16] J., Luo, et al., "Feasibility Study of the Dynamic Network Planning and Management in an End-to-End Reconfiguration (E2R)," *Proc. of WWRF,* Bejing, China, February 2004.
[17] Leonard Kleinrock, Queueing System, John Wiley & Sons, Vol. I&II, 1975.
[18] A. Mihovska, M. Jankiraman, and R. Prasad, "OFDM-MIMO Systems for Fourth Generation: Performance Results," in Proc. of the 7th OFDM Workshop, Hamburg, Germany, September 2003.

Chapter 7

Real-Time Simulation Platform for Cooperative RRM

This Chapter proposes an implementation for the cooperative RRM framework as a real- time simulation platform. The real-time simulation platform is implemented to allow for practical performance evaluation and testing of the proposed in Chapter 2-Chapter 4 RRM framework. The proposed implementation realises the framework as a combination of simulation and testbed contexts. The wireless emulator proposed here is an experimental study of the cooperative RRM mechanisms and how they apply in the context of next generation systems.

The platform supports the inter-working between a next generation RAN and legacy systems (i.e., WLAN, UMTS, GPRS). The platform is based on real-time monitoring of the RANs. The platform demonstrates in real-time application the advantages of the proposed cooperative RRM functionalities for the provision of quality of service (QoS) and congestion management. Another objective for the real-time simulation was to prove the generic nature of the proposed RRM framework. Results are shown in terms of capacity enhancements achievable through use of cooperative RRM in different types of systems (e.g, IMT-A and WLAN) and different deployments of the IMT-Advanced system.

The implementation supports user mobility in a heterogeneous scenario (e.g., inter-system handover), as well as mobility within the RAN (inter-mode handover).

This Chapter describes the practical implementation of the RRM platform and the demonstration set up at a low level. The platform is evaluated for three traffic load scenarios (TLSs) and shows the performance of the RRM framework in a WA and LA deployment for a selected number of services and in terms of handling of higher system loads. The performance is compared to the performance of a WLAN system. Finally, results are shown also in real-time for a high quality video streaming application for a scenario of inter-system handover as a means for congestion management.

This Chapter is organized as follows. Section 7.1 gives the motivation for the implementation. Section 7.2 defines the measurement framework and requirements for the real-time implementation. Section 7.3 describes the functionalities and their interaction for the implementation of the platform components, including split into modules, interfaces, messages exchanged, flow charts and tables. Section 7.4 gives the results and their assessment. Section 7.5 concludes the Chapter.

7.1 Motivation for a Real-Time Simulation

The real-time simulation implementation proposed here was selected as an accurate, convenient and cost-effective solution to evaluate the protocols in real-time.

Simulators cannot reliably represent reality because intrinsically such models are based on problem simplification and abstraction, focusing only on a specific protocol or algorithmic function of great interest, each time. A testbed implementation is an expensive way of performance evaluation that gives a possibility for real-time element operations but some drawbacks are that a testbed can become useless and create misleading results if some specific network conditions and traffic dynamics never occur during a test. Also, any testbed trial is dependent on the operating system that hosts it. If some software modules required for a trial have not been implemented, it may be time wasteful to program all of them. Furthermore, setting up a wireless testbed, in particular for the evaluation of the cooperative RRM proposed here, requires that several software modules related to the wired and the wireless domain are available, in order to ease the integration process.

Although, performance evaluation results were already presented in the previous chapters based on use of simulation tools (e.g., C++, OPNET and OMNET), the real-time simulation platform offers an added value by allowing for evaluation based on actual protocol implementations and applications. The air interface of the reference IMT-A system belongs to a testbed configuration while the BSs, GW and CoopRRM (see Figure 2-1) operations are simulated on a computer that acts like a real entity controlling any other element of the platform. Therefore, the BS, GW and CoopRRM behaviour of the system is said to be emulated rather that simulated. Wireless network emulators have been extensively deployed worldwide [1].

The protocol reference architecture used for the real-time implementation is the one shown in Figure 1-4.

7.2 Measurements and Requirements for the Real-Time Simulation

The objective of the real-time simulation is to emulate a scenario where a UT will be able to initiate a request of a service with the following requirements:

1. Always connected;
2. Best coverage (strongest signal);
3. Best available bandwidth;
4. Best available QoS.

From the network point of view the system should handle all traffic through the proposed in Chapter 2 cooperative RRM mechanisms in order to:

- Decongest an area either as part of a single RAN or as part of an area covered by multiple RANs;
- Help the initiation of a handover.

For example, for a congestion situation, based on the input received from the monitoring sub-network, the main monitoring module and the CoopRRM perform decision-making processes to identify suitable strategies to relief the effects of the congestion. To that, they have available a set of RRM management techniques, (RMTs), which represent the means by which the allocation of resources to the incoming traffic can be arranged in order to optimize resource utilization.

The topology of the proposed real-time RRM implementation is shown in Figure 7-1.

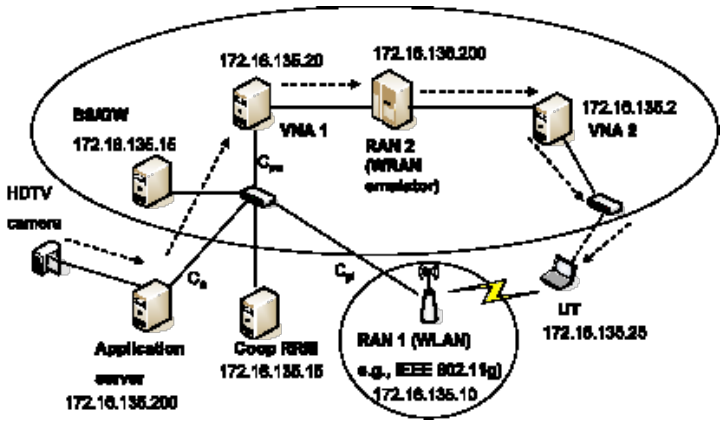


Figure 7-1 Topology of the real-time simulation platform for cooperative RRM.

The platform can function as a stand-alone or as an integrated implementation. In the stand-alone implementation the IMT-A candidate RAN can be emulated by an access point of the type 802.11a/b/g. In the integrated topology (Figure 5-1), the IMT-A candidate RAN is emulated by a testbed configuration. Therefore, the integrated implementation includes additional entities, such as the antennas for the transmission and the reception, a receiver terminal and an entity that controls the transmission antennas at the BS. This last entity is responsible for the management of the transmitters; it gets the measurements from the radio link and sends them to the BS. In this way, the RRM platform knows at any given time the RTTMs of the radio link. The receiver PHY entity is the entity that controls and manages the receiver antenna and it is connected with the RRM platform UT as a network interface card, in order to forward the packets to and from the UT.

The $SRRM_W$ functionality has been implemented in the BS/GW physical entity and, the $SRRM_L$ functionality has been implemented in the CoopRRM physical entity. This was done to simplify the set up for actual demonstrations. In the stand-alone implementation the BS and a GW monitor the state of the system and send alarms and reports to the CoopRRM through the C_{PW} interface. The BS and GW communicate through the C_a interface.

The legacy RAN is emulated in both cases as a WLAN based on the 802.11g wireless standard, and comprises also an $SRRM_L$ module that monitors the system state and informs accordingly the CoopRRM through the C_{PL} interface. The UT is capable of connecting to all the modes of the WRAN and the legacy RAN, using a high-level application that exchanges XML formatted messages with the CoopRRM.

The integrated implementation was used to show results in terms of user – perceived QoS for a real-time high quality video streaming. An HDTV camera captures video and sends it to the application server that streams it to the UT. The UT has two network interfaces, an Ethernet card and a wireless card. When the UT is connected to RAN 2, the Ethernet card is enabled and the wireless is disabled and the opposite happens when the terminal is connected to the legacy network through the access point.

The implementation of the individual modules is proposed in Section 7.3.

7.2.1 System Requirements

The following technical assumptions were made related to the emulation of the IMT-A RAN, to be able to realise a stand-alone implementation for the cooperative RRM

architecture. To be able to emulate the chosen reference IMT-A RAN as an adaptive system operating in the three main scenarios (i.e., LA, MA, and WA), the characteristics were assumed as shown in Table 7-1.

Table 7-1

Parameters for the Emulated Reference RAN

Parameter	**FDD mode (2 x 20 MHz)**	**TDD mode (unpairded 100 MHz)**
Center frequency (GHz)	4.2 (UL), 5.0 (DL)	5.0
Number of subcarriers in OFDM	512	2048
FFT BW (MHz)	20	100
Signal BW (MHz)	16,25	81,25
Number of subcarriers in use	416	1664
Subcarrier spacing (Hz)	39062,5	48828,125
OFDM symbol length (excluding guardtime) (ms)	25,6	20,48
Guardtime / cyclic prefix (ms)	3,2	1,28
Total OFDM symbol length (ms)	28,8	21,76
Chunk length in OFDM symbols	12	5
Chunk duration (ms)	345,6	108,8
Physical chunk size (KHz x ms)	312.5 x 345.6	781.25 x 108.8
Chunk size in symbols	96	80
Duplex guard time or transition gap TX/RX (ms)	-	19,2
OFDM symbols per frame (UL or DL)	12	15
Chunks per sub-frame (UL or DL)	52	312
Frame duration (ms)	691,2	691,2
BCCH duration (ms)		
RAC duration (ms)		
Control super-frame duration	172,8	130,56
Frames per super-frame	8	8
Super-frame duration excluding control (ms)	5,5296	5,5296
Modulation alphabet and coding schemes	Bits per symbol	Coding rate
QPSK 1/2	2	0,5
QPSK 3/4	2	0,75
16QAM 1/2	4	0,5
16QAM 3/4	4	0,75
64QAM 2/3	6	0,67
64QAM 3/4	6	0,75
Raw Bit Rate per Chunk (Kbps)		
Modulation alphabet and coding schemes	FDD mode (2 x 20 MHz)	TDD mode (unpairded 100 MHz)
QPSK 1/2	278	735
QPSK 3/4	417	1103
16QAM 1/2	556	1471
16QAM 3/4	833	2206

64QAM 2/3	1111	2941
64QAM 3/4	1250	3309
Aggregated DL or UL Raw Bit Rate per Frame (Mbps) (DL and UL for 1:1 asymmetry)		
Modulation alphabet and coding schemes	FDD mode (2 x 20 MHz)	TDD mode (unpairded 100 MHz)
QPSK 1/2	14,44	36,11
QPSK 3/4	21,67	54,17
16QAM 1/2	28,89	72,22
16QAM 3/4	43,33	108,33
64QAM 2/3	57,78	144,44
64QAM 3/4	65,00	162,50
Aggregated Raw Bit Rate per Frame (Mbps) (DL and UL for 1:1 asymmetry)		
Modulation alphabet and coding schemes	FDD mode (2 x 20 MHz)	TDD mode (unpairded 100 MHz)
QPSK 1/2	14,44	72,22
QPSK 3/4	21,67	108,33
16QAM 1/2	28,89	144,44
16QAM 3/4	43,33	216,67
64QAM 2/3	57,78	288,89
64QAM 3/4	65,00	325,00

The following system modes were assumed: time division duplex and frequency division duplex (TDD and FDD, respectively) with exemplary raw data rates for TDD of 100 MHz and for FDD of 2x20 MHz. The following calculation metrics were assumed for calculating the BER per chunk, subframe and frames, (see Equations 7-1 to 7-5), respectively:

$$BitRate_{Chunk} = \frac{SymbolsPerChunk \cdot BitsPerSymbol \cdot CodingRate}{ChunkDuration} \quad \textbf{(7-1)}$$

$$BitRate_{SubFrameDLorUL} = \frac{SymbolsPerSubFrame_{DLorUL} \cdot BitsPerSymbol \cdot CodingRate}{FrameDuration} \quad \textbf{(7-2)}$$

$$SymbolsPerSubFrame_{DLorUL} = ChunksPerSubFrame_{DLorUL} \cdot SymbolsPerChunk \quad \textbf{(7-3)}$$

$$BitRate_{Frame} = \frac{SymbolsPerFrame \cdot BitsPerSymbol \cdot CodingRate}{FrameDuration} \quad \textbf{(7-4)}$$

$$SymbolsPerFrame = ChunksPerFrame \cdot SymbolsPerChunk \quad \textbf{(7-5)}$$

The above parameters are used to determine the status of the IMT-A candidate RAN.

To assess the performance of the proposed RRM framework, different user classes, with different service characteristics were identified for the IMT-A candidate system and derived from [2]. The groups of service classes and their characteristics were defined in Chapter 3 (see Table 3-1).

The IMT-A system modes were emulated with an access point into which three modes are imported and controlled by a workstation (i.e., the BS), emulating the IMT-A RAN. An integrated implementation provides an emulator of the IMT-A air interface as a testbed configuration. This was used to assess the user-perceived QoS.

7.2.2 Performance Requirements

The performance measurement is an effective means of scanning the whole network at any time and systematically searching for errors, bottlenecks and suspicious behaviour. Chapter 2 proposed KPI aggregation to deal with the many input and output parameters indicative for the network performance and as a means to assist the RRM decision process with a minimum set of metrics for tracking the system progress towards a performance target [3].

The most important KPIs used in the real-time RRM framework implementation are the *delay,* expressed as the time needed for one packet of data (or a flow) to get from one point to another; the *jitter,* expressed as the delay variation of the received packets (inter-RAN flows) over time; the *peak user data throughput,* expressed as the maximum rate achieved during the transmission of data in the network; and the *mean user data throughput,* expressed as the average rate achieved during the transmission of data in the network. These KPIs were defined in Chapter 2, Section 2.2.1.3, (see Equation 2-2 to Equation 2-14).

Equation 2-4 defined the load L in a generic way as a function of the total capacity. The definition was obtained by a defined dependency between the load L and the delay (τ). If the maximum load at which a system can function without entering a congestion state is given by L_{th}, in a low network load situation, or $L < L_{\text{th}}$, the delay value (τ) can be represented as a typical delay (τ_{typ}). When the load increases and gets in the congestion zone, the delay value then augments very quickly. The formula in Equation 2-4 considered the influence of a congestion threshold parameter (CT) that showed when the congestion zone would be reached. In the scope of the cooperative RRM investigated in the real-time platform implementation, once this critical value has

been reached, the CoopRRM entity will receive a request for handling the arisen congestion situation and an algorithm will be activated. The congestion threshold, *CT,* is the load value *L,* expressed in percentage of the total capacity, chosen to identify a congestion situation, and is used to indicate the upper congestion limit.

The measurements defined below are used in the proposed implementation to detect the status of the system. In addition, these can also be classified as triggers that can necessitate or cause handover, as explained in Chapter 2.

To ensure real-time monitoring functionality for the support of inter- and intra-system cooperation, the following measurements are provided to the SRRM/CoopRRM (see Figure 1-4) as a minimum required information.

A very important measurement is the **received signal strength** in the UT, the **interference level** and the **C/I ratio**. This allows for concluding on the reception quality of the actual configuration and the possibility (or the necessity) of doing a handover to other cell or a RAT. In the IMT-A system, these measurements are based on the UL and DL synchronization pilots and are performed by either the UT, BS/RN, on the IMT-A RAN, but also on the legacy RANs, when necessary. Three different types of measurements should be available intra-frequency, inter-frequency and inter-system, the last one requires a multi-mode UT.

The **transmitted power** of the BS/UT is reported to the SRRM/CoopRRM entites**.** This is just a report of the transmitted power setting in a precise moment. Pathloss measurements can also be measured as the difference between the transmitted power and the received signal strength.

For the execution of the RRM mechanisms related to QoS some quality measurements are also needed. These are a measure for the quality offered and perceived by the UT and GW and to compare it with the required quality. Measurements are performed on the user data flow in order to determine the QoS level and compare it with pre-determined thresholds. The QoS indicators are the block error rate (BLER), the retransmitted block rate or the bit rate at different layers level (e.g., the PHY layer with instantaneous bit rate, MAC layer with throughput or IP layer level). For the IMT-A RAN these are performed by the UT and GW.

For all the RRM mechanisms the **cell load** measurement is common**.** The cell load corresponds to the currently used resources in comparison to the available in the RAN resources, at different levels. The cell load is measured at the PHY layer as the transmitted power or it can be derived from the bit rate, the number of used chunks, etc.

For the legacy RANs the cell load is defined in accordance with the specifics of that RAN (see Chapter 2), but in general, it is considered how the measured load compares to a predefined threshold, or whether $Load < L_{th}$.

The **UT velocity and location** are two needed measurements for the execution of the location-based mechanisms**.** This is important for the RAN/cell selection during handover. As a minimum requirement the system should know to which BS/RAN the UT will be attached and what is the coverage area of the serving BS, a more detailed position determination should be performed by the GW, using the received signal strength measurements or satellite measurements (GPS) .

The KPIs are calculated based on the performed measurements, after performing an aggregation procedure and the outputs are forwarded to the CoopRRM.

The aggregation procedure is based on the reward function defined in Equation 2-1 and consists in applying the function by taking as an input the tuple (KPI_1,…, KPI_k) and producing as an output a real value. From the aggregation the general set of KPIs for each system are obtained as (KPI_{RAN1},…, KPI_{RANr}) and, finally, a general KPI showing the overall system behaviour ($KPI_{overall\text{-}system}$). This is the KPI that would trigger a decision and a corresponding RRM mechanism.

7.2.3 Traffic Load Scenarios

The traffic load scenarios (TLSs) were defined in Chapter 2 in relation to the theoretical assessment of the proposed cooperation mechanisms (i.e., congestion, admission and load control). The TLSs are used in the real-time implementation as categorisation of congestion situations and are based on parameters that introduce load augmentation. The TLSs are also associated with three service sets in the real-time implementation. This approach was adopted so that the TLSs are system-independent, which is important considering the generic nature of the proposed cooperation mechanisms.

The TLSs are used as an indicator of which RRM technique must be selected to resolve an occurred congestion or load situation. This is achieved by associating a *'High-Medium-Low (H-M-L)'* values for KPI parameter as defined in Section 7.2 and indicating the resources availability, the user-perceived QoS and the level of congestion. To determine the system states, the TLSs are represented by a logical tree, where the outcome of the KPIs calculation generates a certain TLS. This is shown in Figure 7-2 where the number of generated states during the TLS is 27 [2].

The KPIs related to QoS (i.e., delay, jitter, etc) require that the RRM technique can improve the QoS state if the final calculation indicates a *'high'* state for the TLS. Each of the states at each level requires the execution of a generic or cooperative RRM technique.

In the proposed real-time implementation, the TLS are generated by the process shown in Figure 7-3. To manage the resources for a given TLS, users are prioritised according to a user and application prioritisation process as specified in Table 3-4 to Table 3-6. By submitting all these values different kinds of traffic were emulated and more freedom was ensured for creating different types of traffic. A *'medium'* or *'high'* state generates an alarm message that triggers a suitable RRM technique. The alarm message is based on the values of the calculated KPIs. This is achieved by the monitoring process.

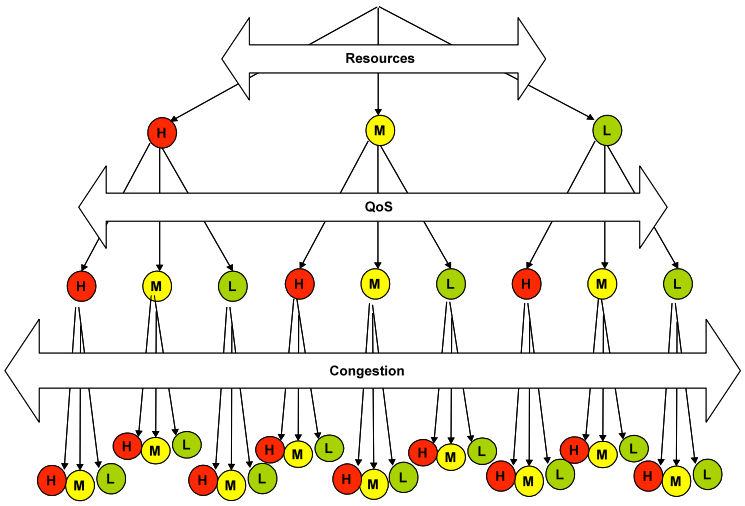


Figure 7-2 Logical tree for TLSs evaluation.

For the evaluation of the real-time RRM platform, three TLSs describing the resources availability in terms of load, congestion, and mean user and data throughput were defined: normal hour (low), bus*y hour (medium)* and emergency (high). The TLS corresponding to one of the three TLSs is translated into the number of users. Different user classes, with different service and radio capabilities were identified for the IMT-A candidate system and derived from [5], (see also Chapter 3). The set of possible services was associated with a given user profile.

The alarm message generated by a *'high'* state has a structure as shown in Figure 7-4. Similar interface architectures can be developed for a query network parameter message or user information message [2]. The relationship between the alarm

generated because of a *'high'* state detected by the monitoring process is described in Figure 7-5.

Each KPI is calculated separately and compared with the given thresholds provided from the operator. An 'alarm' (AL) is created when these values are within the 'red' dots. Within an interval (in Figure 7-5 this time is set to be 20s) the KPIs that show after an initial calculation an 'alarm' value are recalculated. The 'green' dots indicate a relaxed KPI value after the recalculation that does not generate an alarm message. If recalculation still indicates that the KPIs are close or larger than the pre-defined threshold, an alarm message is generated with a structure depending on the type of KPI that has triggered it.

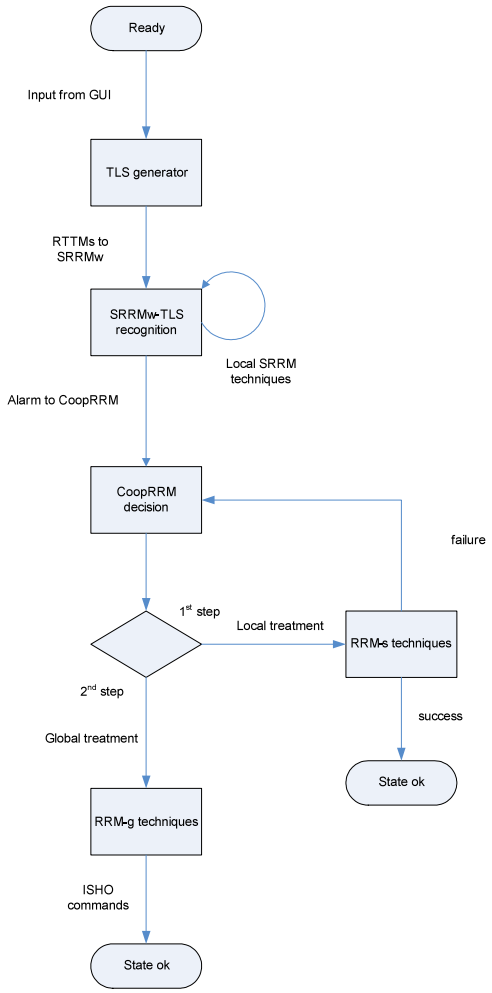


Figure 7-3 TLS generation and selection process.

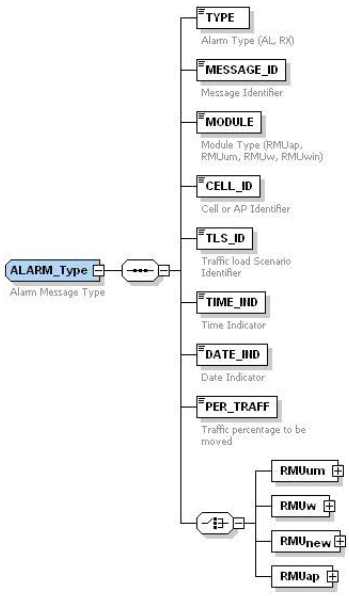


Figure 7-4 Structure of an alarm message.

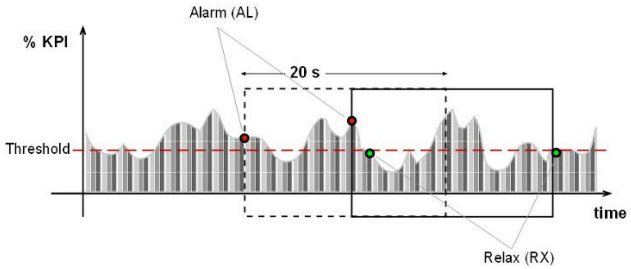


Figure 7-5 Monitoring, calculation of KPIs and a generation of an alarm message.

Calculation of the KPIs is one of the major roles of the monitoring unit. In the proposed implementation the message exchanges between different entities are *xml*-based and are transported over the TCP/IP protocol. This is further explained in the following sections.

7.3 Functionalities of Implemented RRM Modules

The functionalities described for each RRM module are the ones used for the real-time implementation.

7.3.1 CoopRRM Functionalities

It is proposed that the logical functionality of the CoopRRM is divided in a cooperative/generic part (RRM-g) and a specific part (RRM-s) for each RAN with the RRM-g part containing the functionalities common to all RANs. The RRM-g provides a common interface towards upper layer functions/protocols. The RRM-s handles the

specific details of each RAN. The RRM-g is used to handle the UT in order to provide the desired bandwidth and QoS demanded for a service.

The first task of the CoopRRM is to control the network status and alarm information, the second one is the handling of the RRM and the third one is the handling of the on-demand requests from the UT. For example, during inter-system handover (ISHO) from the WRAN to a legacy RAN, the CoopRRM requests from the $SRRM_L$ an approval of the authentication for a specific user before the handover. This requires the following interactions. First, either the $SRRM_W$ or the $SRRM_L$ sends an alarm to the CoopRRM, indicating the RAN status and user information. This information is filtered and upon the result, the CoopRRM executes an RRM-g technique in order to initiate inter-system handover.

The alarm and status information provide the driven force for the handling of the cooperation between the WRAN and the legacy systems. In summary, the CoopRRM implements the following fucntionalities:

- Receive alarms from $SRRM_W$ and $SRRM_L$;
- Demand the status from the monitoring unit;
- Demand status from the SRRM;
- Provide authentication;
- Enable RRM-g.

The internal structure of the CoopRRM is shown in Figure 7-6.

The CoopRRM includes the following modules:

- Service Handler (SHM);
- RRM-g;
- Status Collector (SCM).

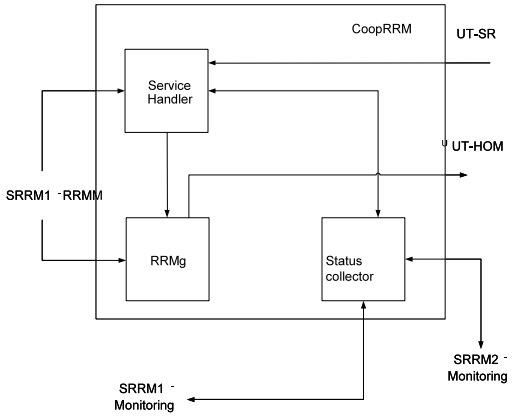


Figure 7-6 Internal structure of the CoopRRM module.

The SHM is in charge to change the modes of the CoopRRM. The modes are three, the first one is the *passive* mode where the CoopRRM waits for any of the signals from the interfaces, the second one is the *monitoring* where the CoopRRM waits by counting down a specific amount of time and the third one is the *active* where there is an ongoing process. The SHM, also creates and handles the cases that are driven from the input of each interface.

7.3.2 $SRRM_W$ Functionalities

The $SRRM_W$ implements the monitoring unit and the RRM-s unit. It accepts information from the GW in a structured format. This information is in a form, which cannot be used directly from the CoopRRM. The calculation of the KPIs is a major functionality of the monitoring unit. Its role is more passive than active, while constantly reading the flow of information from the GW, it calculates the KPIs, and outputs the results to the RRM-s unit or the CoopRRM. The results require that the *monitoring unit* is implemented in two different sub-entities. The first is driven by the monitoring unit, where in every cycle of calculation, a comparison of the results is performed between each KPI and the predefined threshold values. When the value of the calculation is above or lower than the threshold value, (the latter is based on whether the KPI is increased or decreased to the maximum value), then alarm signals are created and a list of the current information is sent to the RRM-s for local management inside WINNER or to CoopRRM for global management. The list contains a summary between the time of calculation which happens in the monitoring unit by providing information on the KPIs, the number of users that are currently connected, the type of service per user and the current mode of the Base Station. The second section of the Monitoring Unit is the on demand request for status driven from the Coop-RRM. The

information exchange is the same as before but the only difference is the type of message that has arrived at the Coop-RRM.

The RRM-s unit performs the specific RRM techniques for handling user requests inside WINNER, for performing intra-mode handovers (IMHO). It also gets alarms which trigger the RRM-s techniques for managing the performance of the network. If the RRM-s techniques are not successful then the alarms are forwarded to the CoopRRM for triggering of the RRM-g techniques. The functionalities of the $SRRM_W$ entity can be summarised as follows:

- Receive real-time traffic measurements;
- Calculate KPIs;
- Forward alarms to CoopRRM;
- Provide status to CoopRRM on demand;
- Enable RRM-s.

The internal structure of the $SRRM_W$ module is shown in Figure 7-7.

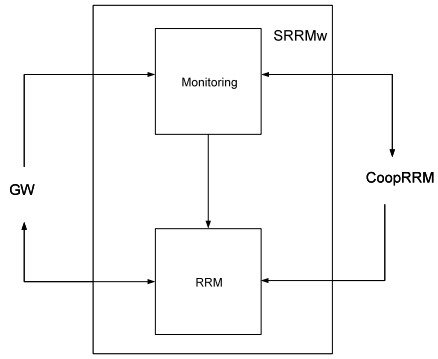


Figure 7-7 Structure of $SRRM_W$.

The required computing power for the $SRRM_W$ is quite high, because of the implementation of the monitoring process and the execution of RRM-s techniques. Therefore, the number of other processes going through this module is downsized to a minimum. The processes are based upon the sequence of monitoring data, process information, and execution and the interfaces C_a and $C_{PW.}$ The $SRRM_W$ receives RTTM reports and stores the values locally. The messages are grouped hierarchically and based on the identification sequence from the BS. The grouping is continuous per BS mode. A hierarchical group provides faster search results for further processing of the messages and for later use based on the demands of the CoopRRM.

7.3.3 $SRRM_L$ Functionalities

The SRRML functionality provides only status information towards the CoopRRM. In order to enable this two-way communication from both sides, an extra interface is required. During a monitoring procedure, the SRRM will forward alarms to the CoopRRM and at the same time the CoopRRM is able to request on demand the status information from the SRRML. The procedure is exactly the same as it is between the CoopRRM and the monitoring unit on the WRAN side. In summary, the SRRML provides the following functionalities:

- Provides status on demand to CoopRRM;
- Forwards alarm to CoopRRM.

The monitoring process for the SRRML is similar to the one described for the SRRMW. In order to fulfill the requirements of calculation of the KPIs, each message is labeled with a unique queue number and time/date information. When a specified amount of messages have been stored locally and the window for processing the results has collected the information, the actual process starts and each KPI is calculated based on formulas. Each KPI is compared with a list of thresholds, specified and provided by the network operator. When the value of a KPI is over or under the limit of a threshold depending on the type of the threshold, the SRRML sends an alarm message describing the congested situation, and providing the hierarchical information of the BS and user information. This message in turn is sent to the CoopRRM. The SRRML communicates through the CPL interface that is TCP/IP based. The internal structure for the SRRML is shown in Figure 7-8.

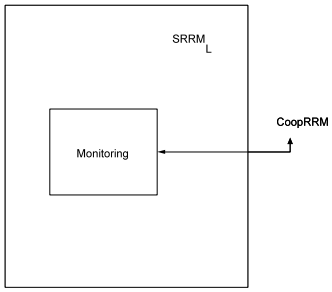


Figure 7-8 Structure of $SRRM_L$.

7.3.4 User Terminal (UT)

The UT is platform-independent, while the only limitations are the hardware capabilities and the number of simultaneously active connections. The capabilities of the UT are divided into two sections. The UT is able to request a handover, which requires communication with the $SRRM_W$ or the CoopRRM. This communication will be done through the GW/RRMserver by sending special messages in a structured XML format. At the same time the UT is able to receive acknowledgements from the $SRRM_W$ or the CoopRRM which is another XML-structured message providing an approval of the request and a summary of information of the handover command. A very important functionality of the UT is the new service request, where the user will request a new service and this message is sent to the GW or the CoopRRM for handling the request. In summary, the UT implements the following fucntionalities:

- Communication with CoopRRM and BS/RRM Server/GW (request for handover)
- Receive handover command;
- Send new service request.

In the practical implementation, the UT will run the following processes:

- **Service request:** this is the process that will make the new service request for the user. The UT will have a graphical user interface (GUI) through which the user can select the desired service. Then this process will recognize the service and it will forward the request to the BS/RRM Server when the UT is connected to a WINNER mode.
- **Network selection:** Network selection is activated as a result of the scanning process, and then computes and evaluates the scanning reports according to the selected service. Subsequently, a decision is taken whether an inter-system or intra-system handover is required. The request is forwarded to the CoopRRM. The internal structure of the UT is shown in Figure 7-9.

The *Service Request Module (SRM)* is the module that requests a new service for an ongoing user. The SRM sends the request for a new service to the CoopRRM entity, with all the necessary user and requested service information.

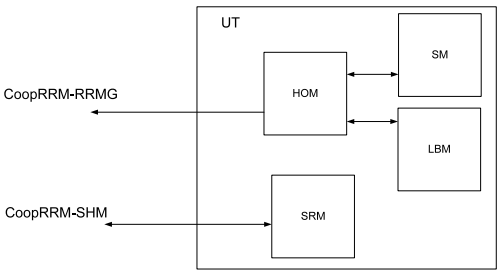


Figure 7-9 Internal structure of the UT.

The *Handover Module (HOM)* is the module that requests an ISHO/IMHO to the CoopRRM or $SRRM_W$ and performs it (UT-initiated handover). The module gets the commands either through the GW (for the IMHO) or directly from the CoopRRM (for the ISHO) and it will perform the handover commands, which change the mode of the WLAN card (a/b/g) in order to perform an IMHO or to change the network (IP, etc.) between the IMT-A system and the WLAN system. The HOM is connected to the *scanning module (SM)* for the support of the UT-initiated handover. The UT will be capable of scanning the spectrum for all the networks that are available. The SM implements a scanning function. The results of the scanning are forwarded to the HOM to decide if handover is necessary. The *Location-Based Module (LBM)* collects location information for the UT.

7.3.5 BS and GW

The BS and GW modules are implemented as one entity. This entity implements the following processes:

- TLS – traffic generator;
- Link controller;
- New user request;
- Mode changer;
- Measurement receiver.

The GW is in charge of the interoperability between the BS and the rest of the RAN elements. Its main task is the decoding of the network specific information. The GW extracts information about RTTMs from the BS on predefined intervals configured by the system. The RTTMs are an indicator for the network parameters and actual

operating modes of the reference WRAN. A selection of RTTMs that would be sent via a dedicated interface are summarized in Table 7-2.

Table 7-2

Summary of RTTMs Obtained from the BS

RTTM	Type
RTTM1	Latency per user
RTTM2	Latency
RTTM3	Erroneous UL packets per user
RTTM4	Total UL packets per user
RTTM5	Erroneous DL packets per user
RTTM6	Total DL packets per user
RTTM7	Erroneous UL packets
RTTM8	Total UL packets
RTTM9	Erroneous DL packets
RTTM10	Total DL packets
RTTM11	Lost UL packets per user
RTTM12	Lost DL packets per user
RTTM13	Lost UL packets
RTTM14	Lost DL packets
RTTM15	Peak throughput per user
RTTM16	Average throughput per user
RTTM17	UL payload data (Kbytes)
RTTM18	DL payload data (Kbytes)

The different TLS are associated to different values of the bandwidth/delay/jitter. This process generates the traffic on the link according to the selected TLS and reports to the link controller, in order to change the state of the BS. The results of the TLS are forwarded to the reporting module that sends the RTTMs to the $SRRM_W$. These RTTMs are also the ones stored locally in the $SRRM_W$.

The messages are grouped hierarchically and based on the identification sequence from the BS. The grouping is continuous per BS mode. A hierarchical group provides faster search results for further processing of the messages and for later use based on the demands of the CoopRRM. Continuously, the process of 'receive' and 'calculate' is straightforward because there is not any major event that would cause an 'H' state, even if the messages are sent asynchronously. In order to fulfill the requirement of calculation and produce the specified KPIs, each message is labeled with a unique queue number and time/date information. When a specified amount of messages have been stored locally and the process window has closed, the actual process of KPI calculation starts.

7.3.6 Signaling Associated with the RRM in the Real-Time Simulation

The required signalling during an inter-system handover is shown in Figure 7-10.

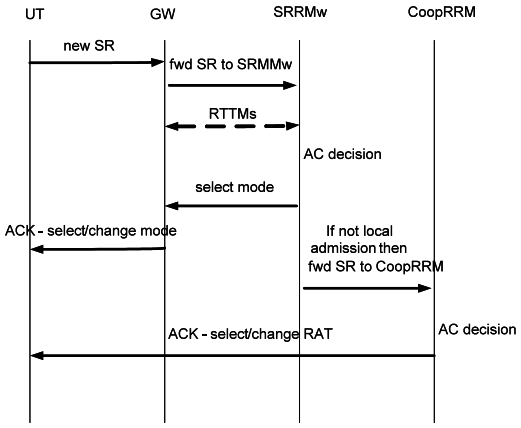


Figure 7-10 Required signaling for the handling of a new service request that initiates an inter-system handover.

The BS/GW receives from the UT the request for the new service and forwards all the necessary information (i.e. user, IP address, location, service, mode, etc.) to the SRRM entity, which knows the status of the network. The BS/GW sends all the time the RTTMs to the $SRRM_W$, which is aware of the status of the RAN modes. The BS/GW gets the decision of the admission (or not) from the $SRRM_W$ entity. This decision will include information about the mode that will serve the user, so the BS/GW selects/changes the mode of the user. The BS/GW sends the ACK to the user about the admission to the network and the information about the required mode.

If the service request cannot be handled inside the IMT-A RAN, then the $SRRM_W$ forwards the request to the CoopRRM entity for ISHO. The GW then receives a command to change the mode of the BS. The ISHO command to the UT is sent from the CoopRRM, directly.

If a congestion situation occurs, the $SRRM_W$ executes the local RRM-s techniques for decongesting the network or performing inter-mode handover (IMHO). This again is based on received RTTMs from the GW. Based on the collected RTTMs the GW has collected from the access point, the $SRRM_W$ calculates the KPIs. If the local techniques are not sufficient to handle the situation, the $SRRM_W$ sends an alarm message to the CoopRRM, for higher level central management (ISHO).

The GW then receives a command to change the mode of the BS. The ISHO command to the UT is sent from the CoopRRM, directly. The signaling for handling a congestion alarm is shown in Figure 7-11.

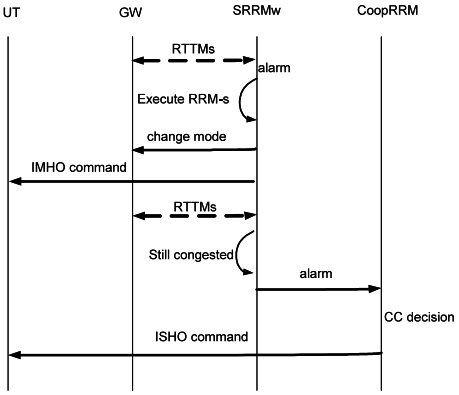


Figure 7-11 Required signaling for the handling of a congestion alarm.

7.4 Results

The real-time simulator was used to assess the cooperative and generic RRM algorithms for three TLSs. In particular, the action of the congestion control mechanism was investigated. The objective is to observe to what extent the proposed RRM framework is effective to handle a congestion situation and what are the approximate congestion thresholds for different system loads. The KPIs related to the load, delay, jitter and throughput were observed in real-time as shown in Figure 7-12. On the left side there are the statistics coming from the BS and then they are computed together with other network statistics and data and the network's KPIs on the right are extracted. The bars on the right are green when the network is in normal condition and are becoming red when there is an overload or critical to overload condition.

The network is congested when the available resources are not sufficient to satisfy the experienced traffic load.

Two congestion scenarios were observed:

1. The network experiments a traffic overload that cannot be totally covered by the available resources, because the traffic rapidly increases inside a group of contiguous cells. This corresponds to the TLS 'sports event';
2. An outage occurs because of unavailability of (part of) the network resources, typically because of malfunctions somewhere.

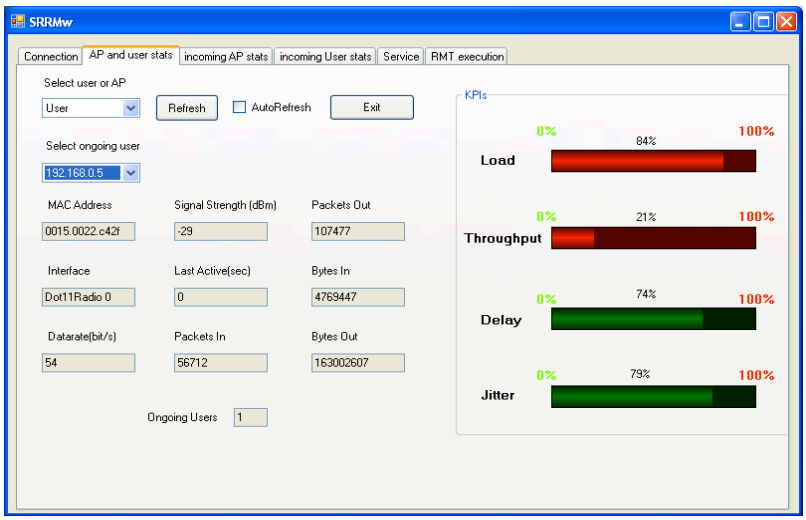
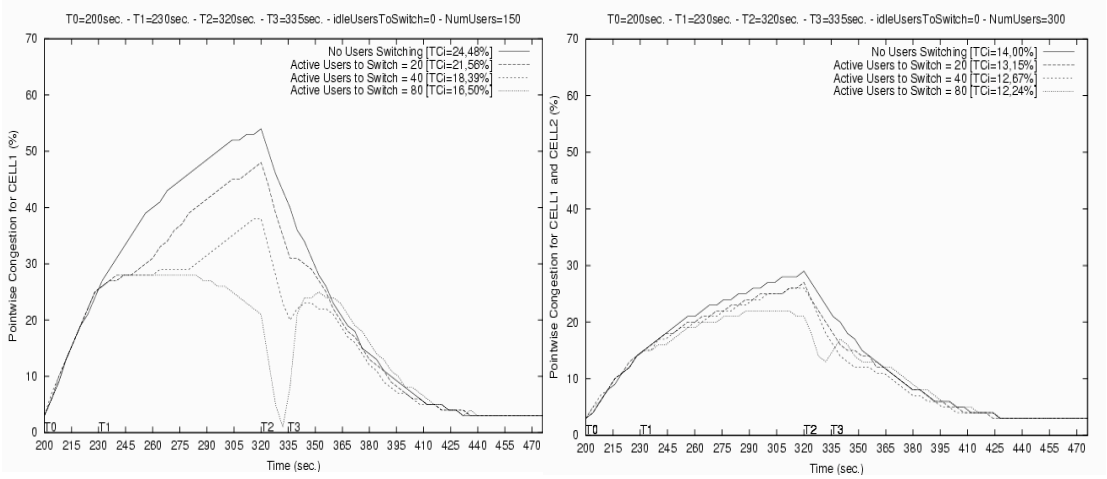


Figure 7-12 Real-time observation of KPIs (in SRRM$_W$) module.

Figure 7-13 a) shows the behaviour of the system for congestion during outage. In such a situation the triggered alarm activates initially a specific or generic RRM mechanism (e.g., intra-system handover).



a) *b)*

Figure 7-13 Congestion caused by outage: a) before RRM and b) after RRM.

In Figure 7-13 the *outageReactionTime* parameter is set to 30 seconds. After this period, some active users will be switched to a different cell and the congestion thresholds will be dropped. This is shown in Figure 7-13 b). In both figures, the congestion has been investigated for different number of users that are forced to perform handover to a different cell.

Based on the load-congestion dependency defined in Equation 2-4, congestion can be detected caused by traffic overload and such a situation would trigger an alarm that would activate a cooperative RRM mechanism.

Figure 7-14 shows the mean user throughput as a function of the congestion threshold for different traffic loads.

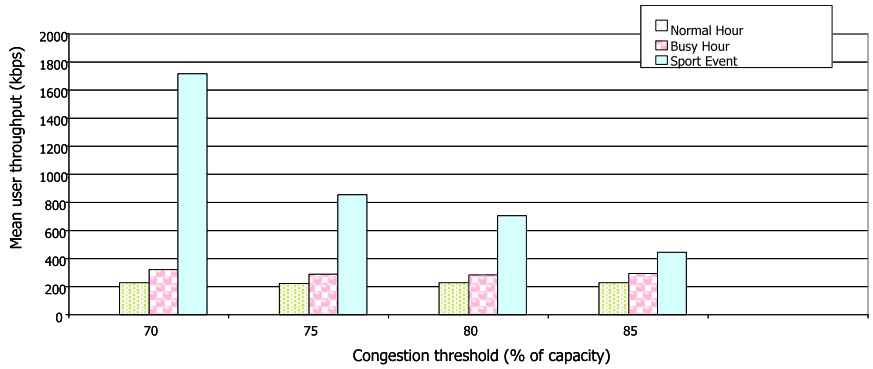


Figure 7-14 Congestion handling as a function of the mean user throughput in the IMT-A RAN (AP operating in LA configuration)

In 'normal' hour the initial number of users requesting connections is low, therefore, the total amount of data transferred is not very high. However, the cooperative RRM algorithms are beneficial for high traffic loads ('sports event') when even for a system operating close to maximum capacity the realizable mean user throughput is still higher than the one for normal hour. The cooperative RRM performs best for a system operating at a 70% of the total capacity, which could be also a suitable value for $L_{\text{th.}}$

A 'busy hour' means that a larger number of users are being connected to the system. Figure 7-15 shows that the cooperative RRM framework allows that more users are getting connected in the busy hour, which is due to the activation of the corresponding load and admission control strategies.

The values of the mean user throughput are also higher because the variety of accessed services is also higher. Figure 7-16 and Figure 7-17 show the efficiency of the RRM framework for two operation configurations, WA and LA, correspondingly.

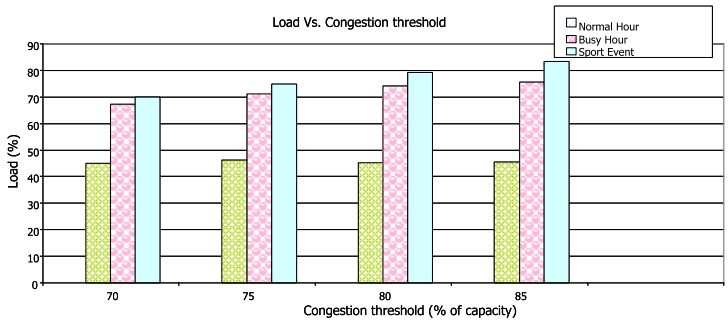


Figure 7-15 Cooperative RRM active during congestion caused by traffic overload in the IMT-A RAN (operating in a WA configuration).

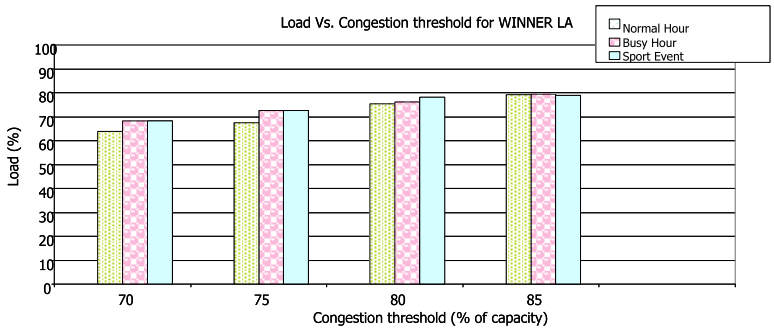


Figure 7-16 Cooperative RRM active during congestion caused by traffic overload in the IMT-A RAN (operating in a LA configuration).

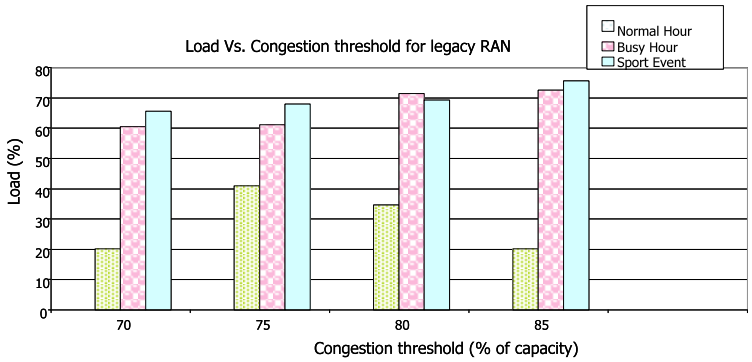


Figure 7-17 Cooperative RRM active during congestion caused by traffic overload in the WLAN.

The differences between the TLS values for a given CT are small, which is due to the higher capacity capabilities of the IMT-A air interface. The cooperative RRM manages to maintain higher loads for longer time before congestion occurs. When $L > L_{th}$, intra-system handover or QoS degradation mechanisms are required as part of the cooperative RRM to bring the system to the normal state. As a comparison Figure 7-17 gives the values for the WLAN.

In the WLAN case, because of the reduced air interface capabilities, the system experiences congestion faster for the same number of users despite the implemented

cooperative RRM framework. However, this shows that the proposed RRM framework is generic in nature and applicable to any type of system. Namely, there is still improvement for high CT values for an operation in 'normal hour'. The gathered RTTMs from both the WRAN and WLAN networks are shown in Figure 7-18.

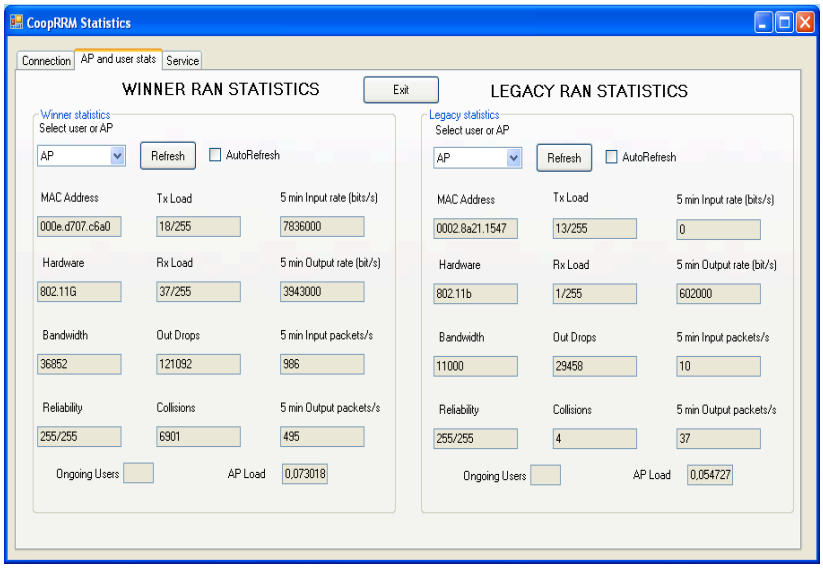


Figure 7-18 RTTM statisitics gatherd for the two systems.

It must be noted that similar results were obtained when inter-system handover was performed [6]. The congestion thresholds were restorable for both cases but within the limits shown in Figure 7-16 and Figure 7-17, respectively. In both cases, handover was performed rather fast (within 20 ms). However, because of the specific RRM for each system the QoS restorable for the legacy RAN users would be lower than for the IMT-A candidate system. The user perceived QoS in real-time is shown in Figure 7-19 for the IMT-A RAN and in Figure 7-20 for the WLAN (after a inter-system handover).

When the user is connected to the WLAN there is a loss in data throughput of about 6-8 Mbps, which is due to the lower capabilities of the air interface, which cannot be resolved by the cooperative RRM framework. The high quality video that was transmitted needed more than 29 Mbps throughput. The WLAN that was used had a maximum throughput of 22 Mbps with the best conditions (maximum transmit power, no collisions, no interferences, very small distance between the access point and the receiver, etc).

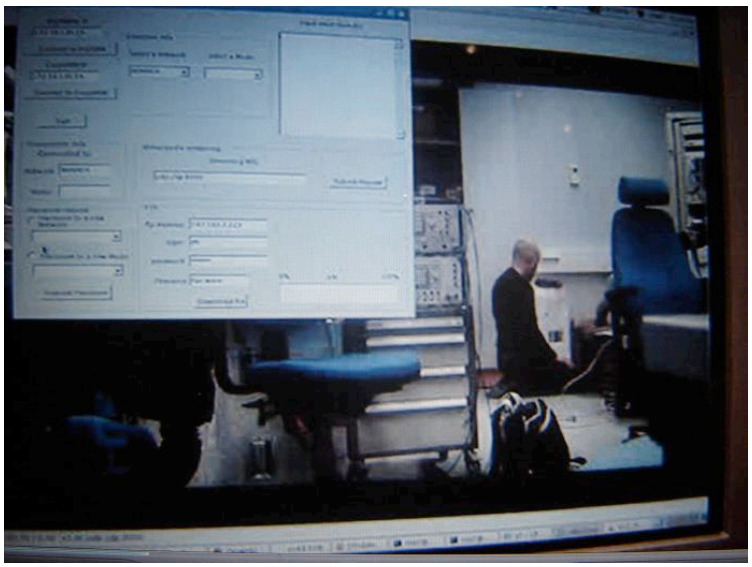


Figure 7-19 Quality of a real-time video streaming application through the IMT-A RAN.

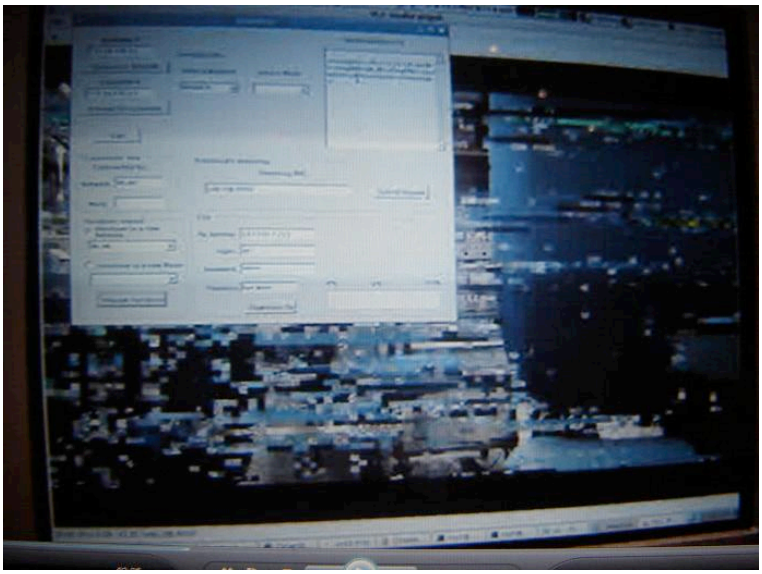


Figure 7-20 Quality of the real-time video streaming application through the WLAN (after inter-system handover).

The IMT-A RAN had a maximum throughput of 100 Mbps. The WLAN is not capable of meeting the throughput requirements of that high quality video, which results in loss of packets, delays and many collisions in the access point, therefore, an inter-system handover to a RAN with lower throughput would result in a reduced QoS perceived by the user.

7.5 Conclusions

The work described in this Chapter proposed a proof-of-concept for the validation of cooperative, generic and specific RRM mechanisms proposed in the context of next generation systems.

It was shown that a new type of RRM framework is needed for an IMT-A candidate systems to address the specifics of a distributed and flat RAN architecture and enhanced air interface capabilities. Further, it is required that RRM mechanisms in the context of next generation systems must be generic related to inter-system interworking. Intra-system interworking can be based on generic and specific RRM mechanisms benefiting from a combined distributed and centralised approach.

With the introduction and integration of several systems with several modes and several layers, resource management becomes a more and more complicated task. Handover and load sharing algorithms must not only maintain the connection at a reasonable quality, they should also consider whether it would be beneficial to move the connection to another system/layer/mode. This decision is not solely based on changing radio propagation, anymore, but also on system load, operator priorities and service quality parameters.

The proposed implementation provides a basis for further enhancements in terms of multiple legacy RANs and real-time traffic generation.

The real-time simulation platform is capable of the following:

- Show the difference in the performance between the IMT-A candidate system and the best available legacy system (i.e., WLAN);
- Show an improved QoS that the IMT-A candidate system provides to the users;
- Emulate the cooperation architecture for the cooperation between heterogeneous RANs;
- Implement and evaluate the cooperation mechanisms;
- Show that the cooperation between an IMT-A candidate systems and legacy RANs is feasible;
- Implement and evaluate the cooperation between transmission modes of the same RAN.

The implementation can be used for generating results in terms of optimised measurements and triggers for handover, as well as in terms of improved throughput

(i.e., reduced delays). The most important feature of the implementation is its generic nature, allowing for the inclusion of entities both horizontally (same architectural level, e.g., routing, intelligent modules) as well as vertically (e.g., implementation of a backbone network with QoS mechanisms).

As part of the future work, the trial platform will include a multiple system/operator scenario, where resources are shared between two operators and advanced spectrum sharing techniques are considered for the IMT-Advanced candidate system. The goal is to assess the framework in terms of additional capacity enhancements based on the resource release at lower layers. This would allow to investigate the effect on the congestion when suddenly more resources are released from spectrum aggregation. Future work plans to investigate the proposed RRM framework and its behaviour for this particular case.

To realize mobility based on Mobile IPv6, the original implementation must be split in three different IP domains. However, this requires the implementation of an additional entity that has routing functionalities and three different network interfaces for routing the traffic between the network domains. It is presumed that the implementation does not require particular changes in the RRM-s or RRM-g mechanisms. Rather, it would give the advantage of taking into consideration the available transport resources in the decision making process. This should be a part of the decision for admission control during inter-system handover. This, however, together with the jitter that can occur during handover because of of packets arriving pout of order, would change the delay curve, and might affect the load/delay dependence used in the proposed validation.

Further, when using Mobile IP, jitter in accordance with the order of arrival of the received packets during handover can occur, which will be visible as peaks in the delay curve. It is presumed that the jitter from Mobile IPv6 will give slightly different results for the load/delay dependence.

References:

[1] P., Zheng and L., M., Ni, "EMWIN: Emulating a Mobile Wireless Network using a Wired Network," in *Proc. of the Fifth ACM International Workshop on Wireless Mobile Multimedia,* 2002, Atlanta, Georgia, USA , Pages: 64 - 71. ISBN:1-58113-474-6.

[2] A., Mihovska, et al., "QoS Management in Heterogeneous Environments," *Proc. of ISWS'05,* Aalborg, Denmark, September 2005.

[3] G., Gomez and R., Sanchez, *End to End Quality of Service,* John Wiley & Sons, 2006.

[4] S., Nousiainen, S., Kyriazakos, et al, "Measurements and Performance Evaluation," Deliverable 2.3, IST project CAUITION, May 2003.

[5] A. Mihovska, et al., "Algorithms for QoS Management in Heterogeneous Environments," Proc. of WPMC'06, San Diego, California, September 2006.

[6] A. Mihovska, et al., "Practical Implementation of Cooperative Radio Resource Management," in *Proc. of the ATSMA NAEC 2008,* Riva del Garda, Italy, September 2008.

Chapter 8

Conclusions and Future Work

This Chapter summarizes the contributions of each chapter and proposes the follow up work that builds upon the achieved results towards the further development of the RRM framework.

This thesis investigated the fundamental benefits of a novel approach to RRM to ensure the interworking between IMT-A candidate systems and legacy systems, as well as the interworking within the reference IMT-A candidate system in support of user mobility and QoS. The main achievement of the performed work is a qualified basic RRM concept that includes the latest advancements in the area of radio access technologies and is based on three types of RRM mechanisms: cooperative, generic and specific.

The proposed RRM mechanisms are used to ensure an appropriate system/RAT selection that guarantees QoS to users as well as efficient network management. The mechanisms are based on a set of selection criteria. This thesis considered the load, the mean user throughput, the distance to the BS, the signal strength and the type of services as the main selection criteria for executing an RRM mechanism.

The proposed general concept for RRM in support of inter- and intra-system interworking and applicable to next generation radio access systems is based on a three-layer RRM framework, comprising *cooperative, generic and specific* RRM mechanisms. The developed concept operates at L2 and L3 of the protocol stack. In particular, *cooperative* and *generic* RRM mechanisms were investigated in relationship to the inter- and intra-system interworking and with the objective to demonstrate the benefits of the proposed RRM framework. Detailed investigation of *specific* RRM mechanisms was not in the scope of this thesis. The performed work included also an experimental set up for cooperative and generic RRM as part of the investigation.

It was shown that a combined centralised and distributed approach to RRM provides scalability and flexibility of the proposed RRM framework. Further, this is an

optimal approach for next generation radio access systems, which would be foreseen of a distributed and flat RAN architecture.

Distributed RRM was proposed for intra-system interworking in situations of low-to-medium loads as a way to shorten the response time to service requests and system performance handlings. In situations of medium-to-high load a centralised approach will be more beneficial even if this means an increased signalling overhead. A centralised entity will help to balance the loads and maintain the system stabilities.

Inter-system interworking requires a centralised approach. It was proposed that the control as a principle is performed by an entity located outside the RANs to maintain the generic character of the proposed RRM framework and to maintain the original RAN architecture of the legacy systems. Inter-system interworking relies on *cooperative RRM*. New systems can benefit by the combined distributed/centralised RRM approach, which allows that the cooperative RRM functions are implemented at lower layers and closer to the air interface (i.e., located in the optional RRM server).

It was shown that significant capacity enhancements, expressed in terms of achievable loads, and number of connected, blocked and dropped users are achievable through the proposed RRM framework compared to the case when the proposed mechanisms were not implemented. Further, it was shown that there is an important dependency between the performance of the proposed RRM algorithms and the triggers for their execution.

Further improvements can be achieved by adding more accuracy to the decision-making process. An approach based on use of fuzzy logic was proposed for improving the decision process during inter-system handover.

The combined centralised and distributed RRM approach can be investigated further in terms of achievable trunking gains from spectrum aggregation. During spectrum aggregation, different parts of the spectrum can be shared and aggregated dynamically to utilize the spectrum as efficiently and fairly as possible. Results reported in literature have shown improvement in information throughput of around 200% in a bandwidth limited network. Spectrum aggregation has been adopted as a way to reuse spectrum that is currently allocated to second and third generation wireless communication systems and to allow IMT-A system to operate in multiple bands. Those systems can use the multiple bands for balancing the load of the networks or for providing required QoS levels. It is also predicted that some of these bands, might be dedicated to specific services or operators, and that other bands, might be shared between different operators and/or different services (e.g. mobile communications and

fixed satellite services (FSS)). In this context, the future research proposed here considers a multiple system/operator scenario, where resources are shared between two systems/operators and advanced spectrum sharing techniques are considered for the IMT-Advanced candidate system. Spectrum aggregation would give a resource release at lower layers. The proposed in this thesis RRM framework allows for resource release at L2 and L3. The achievable gains will be assessed for different scenarios and a trade off function will be established for each of them. In some cases, sufficient congestion/load release can be achieved only by use of spectrum aggregation. The research will determine a cost function that will consider the hardware restrictions, added signaling overheads and delays. This function will be added to the decision process for scenarios requiring cooperative, generic and specific RRM for an optimized approach.

It was shown that choosing the correct trigger for a given scenario can bring improvements in performance of about 10% and vice versa. Choosing a trigger threshold accurately, results in significant decrease of unnecessary handovers. Use of location information, can further enhance the triggering of an RRM mechanism and thus improve system performance. The choice of a trigger and the accuracy in determining which trigger should activate the execution of a certain RRM mechanism can be improved by use of navigation technologies and by introducing computational intelligence to the decision process.

It was shown that the proposed RRM framework allows for policy-based RRM mechanisms as an additional improvement of system performance. Policies were proposed for RAT association and user context transfer in support of mobility management (intra-system handover). The policies can be applied also to inter-system mobility management.

Group or *individual* differentiation was proposed for RAT association in a scenario of a WA BS overlapping a LA BS. Allowing the UT to connect to the BS based on the likelihood that a handover to the WA would be performed because of mobility and service characteristics can decrease the number of unnecessary simultaneous connections. Future work includes an implementation of the proposed strategies into a simulation tool and comparison of the achieved results in terms of achievable resource release (i.e., decreasing the number of unnecessary handovers and simultaneous connections), especially in the context of multi-mode terminals. Further, positioning technologies will be included into the model to more accurately determine the position of each user terminal. The proposed differentiation strategies offer benefits

in the context of self-management of BSs. These benefits also will be investigated within the scope of future work.

Context transfer has advantages and disadvantages, however, based on the network architecture, it was shown that new network functions can be introduced to enable algorithm-diversity that can provide user-centric and service-centric QoS provisioning using policy enforcement. As a first step to achieving diversity, SDU and PDU RLC user context transfer was proposed for radio and IP handover in support of TCP performance. The proposed policies were investigated for different amounts of data to be transferred, different link delays and different polling times. It was shown that the size of the SDU and PDU packets can influence the delay and provide for improved L2 support for high data rates.

Therefore, it would be beneficial to investigate further optimisation policies of the RLC headers to achieve a flexible SDU and PDU size. Context transfer then can be investigated for different SDU and PDU sizes depending on the service requirements and amount of data to be transferred. These investigations would link the SDU and PDU RLC packets size to the particular service context.

The role of the type of I_{bb} and I_g interface in the scope of the proposed policy during user context transfer for radio and IP handover will also be investigated as part of future work as a further study of the interactions between RRM entities within the RAN. For example, inter GW-BS interworking considers both traffic behaviour and the system capacity (i.e., inter-system, inter-mode and intra-mode RRM). The I_{bb} is a multi-to-multi interface. The interface supports distributed RRM functions, such as active mode mobility, interference management schemes, and other distributed inter-BS control and negotiation functions (for example, load balancing). Several spectrum functions are located in the BS that would potentially need this interface.

Communication between BSs in the scope of the IMT-Advanced systems is important because the IMT-A RAN architecture allows that one BS controls another one. Added to an operation in the high frequency range (e.g., 5 GHz) this means that over-the-air-interface for BS-BS communication will not be reliable for non line-of-sight situations. Future research would seek the trade off between the type of interface for a given RRM scenario.

The proposed RRM framework was investigated further in the context of load and admission control. A multi-stage admission control mechanism was proposed as a way to decrease the response times to service requests while balancing the load among the BSs. The proposed multi-stage admission control is particularly beneficial to a

multi-hop communication system, where the coverage is extended based on use of relays. A complete implementation of the proposed algorithm will allow for performance assessment and optimisation of the strategies in different situations. Balanced loads lead to an improvement of the SINR distribution of about 30%.

Follow up work utilizes the enhancements achieved through the proposed intra-system RRM mechanisms (at L2 and L3) in combination with enhancements at lower layers. The wireless link exhibits a time-varying quality due to fading, shadowing, in addition to multi-user interference. To cope with this variability of the wireless channel, techniques have been developed either at the physical layer or the MAC layer. The main shortcoming of the strictly based PHY layer approaches is that they do not take into account the impact on the upper layers, while the main shortcoming of strict layering MAC layer based schemes is that they are based on "hard" channels, i.e. they use very limited information from the physical layer. This implies that the throughput achievable at the upper layers is only a small fraction of the capacity offered by the PHY layer.

As a first step, a joint time-frequency resource allocation on the DL of an OFDMA system is investigated with the proposed multi-stage admission control. Focus is on decreasing the blocking and dropping probabilities while satisfying a larger number of user requests (i.e., QoS for connected users is not degraded). The proposed research will include as a second step the investigation of strategies such as computational intelligence for the calculation of a priority function. As a third-step the research will focus on a combination of allocation and degradation strategies, including cell selection.

An experimental set up combined the proposed inter-and intra-system RRM mechanisms to demonstrate the benefits of the proposed RRM framework in terms of user-perceived QoS. This was shown for a real-time high quality video streaming.

Further, capacity enhancements were shown in terms of mean user throughput and congestion management with use of the proposed RRM framework. Future work will implement computational intelligence and introduce a number set of real-time access technologies to observe the performance of the proposed RRM framework.

List of Peer-Reviewed Publications:

Book Chapters:

*[1] **A. Mihovska,** J. Luo, E. Mino and E. Tragos: Chapter on "Cooperative Radio Resource Management for Heterogeneous Networks" to be published in the book on *Cooperative Wireless Communications,* to be published by Auerbach Publications (AU#6469), CRC Press, Taylor&Francis Group in 4th quarter of 2008, *Editors Yan Zhang, Hsiao-Hwa Chen and Mohsen Guizani*

*[2] E. Mino, J. Luo, E. Tragos, and **A. Mihovska:** Chapter on "Radio Resource Control and System Level Functions," in the book on *Advanced Radio Technologies and Concepts for Future Mobile Communications* to be published by Wiley Publishers in end of 2008, *Editors: Rahim Tafazolli, Afif Osseiran, Martin Doettling*

*[3] R. Prasad, O. M. Lauridsen, and A. Mihovska: Chapter on "Convergence of NAVCOM towards 4G," in AEROSPACE TECHNOLOGIES AND APPLICATIONS FOR DUAL USE ? A NEW WORLD OF DEFENCE AND COMMERCIAL IN 21st CENTURY SECURITY River Publishers 2008, *Editors: Marina Ruggieri*

Journal Papers:

*[1] **A. Mihovska,** F. Platbrood, G. Karetsos, S. Kyriazakos, R. van Muijen, R. Guarneri, and J. M. Pereira, "Towards the Wireless 2010 Vision: A Technology Roadmap," in Special Issue on Advances in Wireless Communications of the Springer International Journal on Wireless Communications, DOI: 10.1007/s11277-006-9180-0, September 2006.

*[2] **A. Mihovska** and R. Prasad, "Secure Personal Networks for IMT-Advanced Connectivity," *Special Issue of the Springer International Journal on Wireless Communications,* DOI: 10.1007/s11277-008-9485-2, April 2008.

*[3] E. Tragos, **A. Mihovska,** E. Mino, P. Karamolegkos, P. Vlacheas, J. Luo, "Access Selection and Mobility Management in a Beyond 3G RAN," *accepted for publication in the Springer Journal on Telecommunication Systems,* to appear in autumn 2008.

*[4] **A. Mihovska,** J. Luo, E. Mino, E. Tragos, C. Mensing, R. Fracchia**,** "Requirements and Algorithms for Cooperation of Heterogeneous Networks," *in Springer*

International Journal on Wireless Personal Communications, DOI: 10.1007/s11277-008-9586-y, August 2008.

In review:

[5] F. Meucci, **A. Mihovska,** R. Prasad, „OFDMA Multi-User Scheduler with Diversity and Spatial Multiplexing MIMO Schemes," *submitted to IEEE Letters on Communication in* July 2008.

*[6] F. Meucci, **A. Mihovska,** B. Anggorojati, N. Prasad, "Achieving Fairness in the Call Admission Control in OFDMA Systems using a Cross-Layer Approach," *submitted to Special Issue on Fairness in Radio Resource Management for Wireless Networks of the EURASIP Journal on Wireless Communications and Networking,* July 2008.

*[7] A. Osseiran, E. Hardouin, M. Boldi, I. Cosovic, K. Gosse, A. Gouraud, J. Luo, J. F. Monserrat, T. Svensson, A. Tölli, **A. Mihovska,** S. Redana, M. Werner and W. Mohr, "The Road to IMT-Advanced Communication Systems: State-of-the-Art and Innovation Areas Addressed by the WINNER+ Project," *submitted to Special Issue on Next Generation 3GPP Technologies of the IEEE Communications Magazine,* September 2008, expected publication date April 2009.

Peer-Reviewed Conferences:

2001

[1] A. Mihovska, J., Pereira, and R. Prasad, "Wireless Multimedia: Trends and Requirements," in *Proc. of IEEE VTC'01 Spring,* Rhodos, Greece, May 2001.

[2] A. Mihovska and R. Prasad, "Performance Investigation of a Wireless IPv6-based Architecture for Mobile Multimedia Applications," *in Proc of IEEE VTC'01 Fall,* October 2001, Atlantic City, NJ.

2002

***[3] A. Mihovska ,** C. Wijting, S. Ponnekanti, M. Nakamura, and R. Prasad, **"**A Novel Flexible Technology for Intelligent Base Station Architecture Support for 4G Systems**,"** *In Proc Of WPMC'02,* October 2002, Honolulu, Hawaii.

[4] A. Mihovska, M. Jankiraman, R. Prasad, **"**OFDM-Based Wireless Packet Data Transmission at 5 GHz for Future-Generation Systems**,"** *in Proc of the OFDM Workshop 2002,* September 2002, Hamburg, Germany.

2003

[5] A. Mihovska, M. Jankiraman, and R. Prasad, **"**OFDM-MIMO Systems for Fourth Generation: Performance Results," *in Proc of the OFDM Workshop 2003,* September 2003, Hamburg, Germany.

***[6]** S. Ponnekanti, C. Wijting, Y. Awad, M. Nakamura, **A. Mihovska, and** J.M. Pereira, "Flexible Cross-Layer Radio Access Design for Systems Beyond 3G," *in Proc. of IST Mobile Summit 2003,* June 2003, Aveiro, Portugal.'

***[7] A. Mihovska,** S. Ponnekanti, and R. Prasad, "Ensuring End-to-End QoS Through Dynamically Adaptive RRM Techniques," *In Proc Of WPMC'03,* October 2003, Yokosuka, Japan.

***[8] S., Kyriazakos,** A., Mihovska, and J. M., Pereira, "Adaptability Issues in Reconfigurable Environments", *in Proc of IST Proc of ANWIRE workshop* on Reconfigurability, Mykonos, Greece, September 2003.

***[9] A., Mihovska;** H., Laitinen, P., Eggers, "Location and Time Aware Multi-System Mobile Network," *in Proc. of Mobile Location Workshop'03,* Aalborg, Denmark, May 2003.

2004

***[10]** V. Sdraia, E. Mino, M. Lott, **A. Mihovska,** D. Lugara, E. Tragos, S. Ponnekanti, G. Karetsos, "Cooperation of Radio Access Networks: The IST FP6 WINNER project approach," in *Proc of Wireless World Research Forum, WWRF Nr 11,* Oslo, Norge, June 2004.

***[11] A. Mihovska,** G. Karetsos, S. Ponnekanti, "RRM Techniques for Heterogenous Wireless Systems," *in Proc. of Mobile Venue Workshop,* May 2004, Athens, Greece.

***[12]** P., Gelpi, **A., Mihovska ,** A., Lazanakis , G., Karetsos, B., Hunt, J., Henriksson , P., Oillikainen, and L., Moretti, "Scenarios from the WINNER Project: Process and Initial Results," *in Proc. Wireless World Research Forum (WWRF), 11th meeting,* Oslo, Norway, June 2004.

2005

***[13]** M. Lott, V. Sdralia, M. Pischella, D. Lugara, **A. Mihovska,** S. Ponnekanti, E. Tragos, E. Mino, **"**Cooperation Mechanisms for Efficient Resource Management between 4G and legacy RANs," in *Proc of 13th Wireless World Research Forum, (WWRF),* March 2005, Jeju, Korea.

*[14] M., Lott, V. Sdralia, D. Lugara, M. Pischella, **A., Mihovska,** S. Ponnekanti, E. Tragos, E. Mino, "Cooperation of 4G Radio Networks with Legacy Systems," *Proc. of IST Mobile Summit 2005,* Dresden, Germany, June 2005.

*[15] S. Frattasi, **A. Mihovska,** S. Kyriazakos, F. Fiztek, R. Prasad, "Service Architecture and Management in a Beyond 3G System," *in Proc of 14th Wireless World Research Forum, (WWRF),* June 2007, San Diego, USA.

***[16]** **A., Mihovska,** S. Kyriazakos, E. Gkroutsiosis and J. M. Pereira, "QoS Management in Heterogeneous Environments," *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

*[17] **A. Mihovska,** S. Kyriazakos, E. Mino, M. Pischella, E. Tragos, V. Sdralia, "Assessment of RRM Schemes for the Efficient Cooperation of RANs: WINNER Requirements," *Proc. of WPMC'05,* Aalborg Denmark, September 2005.

*[18] **A. Mihovska,** S. Kyriazakos, E. Mino, M. Pischella, E. Tragos, V. Sdralia, "Assessment of Radio Resource Management Schemes for Efficient Cooperation of RANs: An Implementation Approach," *in Proc. of IST EVEREST Workshop,* November 2005, Barcelona, Spain.

*[19] V. Sdralia, E. Mino, M. Pischella, **A. Mihovska,** E. Tragos, S. Kyriazakos, E. Mohyeldin, "Achieving Inter-RAN Cooperation: An Architecture Proposal," *in Proc of 15th Wireless World Research Forum, (WWRF),*December 2005, Paris, France.

2006

***[20]** **A. Mihovska,** S. Kyriazakos, and J. M. Pereira, "Algorithms for QoS Management in Heterogeneous Environments," *Proc. of WPMC'06,* San Diego, California, September 2006.

*[21] P. Karamolegos, E. Tragos, A. Lazanakis, **A. Mihovska,** S. Kyriazakos, G. Champion, J. Lara, L. Moretti, "A Methodology for User Requirements Definition in the Wireless World," *in Proc of IST Mobile Summit 2006,* June 2006, Mykonos, Greece.

2007

*[22] **A. Mihovska,** J. Luo, E. Mino, E. Tragos, C. Mensing, G. Vivier, R. Fracchia, "Policy-Based Mobility Management for Next generation Systems," *Proc. of IST Mobile Summit 2007,* Budapest, Hungary, July 2007.

*[23] E., Mino, J. Luo, R. Fracchia, G. Vivier, **A., Mihovska,** E. Tragos, "Scalable and Hybrid Radio Resource Management for Future Wireless Networks," in *Proc. of IST Mobile Summit 2007,* Budapest, Hungary, July 2007.

[24] **A. Mihovska,** and N. Prasad, "Adaptive Security Architecture based on EC-MQV Algorithm in Personal Network (PN)," *in Proc. of the Second International Workshop on Personalized Networks collocated with MOBIQUITOUS 2007,* August 2007, Philadeplhia, USA.

*[25] E., Tragos, **A., Mihovska,** E., Mino, J., Luo, R. Fracchia, G. Vivier, X., Xue, "Hybrid RRM Architecture for Future Wireless Networks," in *Proc. of PIMRC 2007,* September 2007, Athens, Greece.

*[26] E. Tragos, **A. Mihovska,** E. Mino, and P. Karamolegos, "Performance Analysis of Access Selection and Mobility Management in a Beyond 3G RAN," *in Proc of the 5th ACM International Workshop on Mobility Management and Wireless Access Protocols (MobiWac 2007),* October 2007, Crete, Greece.

*[27] **A. Mihovska,** S. Kyriazakos, and N. Prasad, "A Cognitive Approach to Network Monitoring in Heterogeneous Environments," *in Proc. of WPMC'07,December 2007,* Jaipur, India.

2008

***[28]** A., Klockar, **A., Mihovska,** J. Luo, E. Mino, "Network-Controlled Mobility Management with Policy Enforcement towards IMT-A," in *Proc. of IEEE International Conference on Communications, Circuits and Systems,* May 25-28, 2008, Xiamen, China.

*[29] **A. Mihovska,** J. Luo, B. Anggorojati, S. Kyriazakos, N. Prasad, "Multi-Stage Admission Control for Load Balancing," *in Proc. of WPMC'08,* Sept 8-11, 2008, Lapland, Finland.

*[30] F. Meucci, **A. Mihovska,** B. Anggorojati, and N. Prasad, "Joint Resource Allocation and Admisison Control for an OFDMA-Based System," *in Proc. of WPMC'08,* Sept 8-11, 2008, Lapland, Finland.

*[1][31] **A., Mihovska,** E., Tragos, S., Kyriazakos, P., Anggraeni, N., Prasad, "A Practical Implementation of Cooperative Radio Resource Management," *in Proc. of the ATSMA International Networking and Electronic Commerce Research Conference (NAEC 2008),* September 25-28, 2008, in Riva del Garda, Italy.

[32] O. Cabral, F. J. Velez, **A. Mihovska,** and N. R. Prasad, "Optimization of Multi-Service IEEE802.11e Block Acknowledgement," *to be published in Proc. of the IEEE Radio and Wireless Symposium (RAWCON),* to be held in January 2009, in San Diego, California

Contributions matrix:

#contribution / Chapter	Books	Journals	Conferences
Chapter 1	[3]	[1], [2],	[3], [11], [10], [16], [19], [15], [12], [13], [14], [6]
Chapter 2	[1],	[2], [4]	[7], [9], [16], [14] [20], [22], [27] [8],
Chapter 3	[1],	[4], [3]	[26], [25], [23]
Chapter 4	[1]	[4]	[17], [18], [20], [21],
Chapter 5	[2]	[4], [3]	[22], [28], [26], [22]
Chapter 6	[1]		[29]
Chapter 7	[1]		[17], [18], [20],[31], [27], [21]
Chapter 8		[5], [6], [7]	[30]

[1] * Publications related to this thesis